Prioritizing Observational Associations With Family History of Disease and Mendelian Randomization

Dissertation Advisor: Chirag J Patel Danielle Rasooly

Prioritizing Observational Associations with Family History of Disease and Mendelian Randomization

Abstract

Elucidating the biological relevance of risk factors of chronic multifactorial disease is indispensable for identifying new putative intervention targets for disease prevention and therapeutic development. However, discovering novel associations in chronic disease and inferring causality is a challenge. There remains an unmet need for identifying associations between modifiable factors and phenotypes to uncover biological relevance in presence of confounding or reverse causality. One such tool is family history of disease, which reflects the effects of genetics and modifiable factors, including environment, and behavioral factors, and is a major risk factor for chronic disease, including type 2 diabetes and cardiovascular disease. Another tool is Mendelian randomization (MR), a statistical technique that leverages genetic variation for uncovering causal relationships between potentially modifiable exposures and outcomes. We propose to leverage family history of disease and MR to prioritize observational associations in chronic disease, thereby identifying potentially novel associations that may help elucidate the underlying biological (genetic and/or environmental) mechanisms by which risk factors cause disease.

Contents

Acknowledgements

This body of work would not have been possible without the mentorship, guidance, and endless support of many individuals. First and foremost, I would like to express my sincerest gratitude to my dissertation advisor, Dr. Chirag Patel, for the privileging opportunity to work with him and to be a member of his team, and for his continual support and encouragement, contagious excitement for our projects, and unwavering enthusiasm for our potential as academic scientists. This thesis would not have been possible without his endless support and encouragement. Second, I am extremely grateful to my collaborators, Dr. Muin Khoury, Dr. John Ioannidis, and Dr. Arjun Manrai, for their partnership in our projects, thoughtful insight on our findings, and generous mentorship throughout the years. I would also like to express my sincere appreciation to my dissertation committee, Dr. Isaac Kohane, Dr. Jose Florez, and Dr. Ronald Kahn, for their continual advice, insight, and guidance throughout writing my dissertation. To the wonderful faculty and staff at Harvard Medical School Department of Biomedical Informatics (DBMI), thank you for making DBMI a wonderful new home for me. To the incredible team members of the RagGroup, thank you for your ongoing support, thoughtful discussions, and collaborations. And most of all, I thank my family and close friends for their support and motivation that has kept me driven and propelled. I dedicate this thesis to my parents for their limitless support and unwavering encouragement for my pursuit of science, and to my husband, for being my inspiration.

0
Introduction

Family health history provides a gateway to investigating the complex interplay of genetic influences, shared environment, lifestyle, and behavioral factors known to increase risk for many common chronic diseases [1,2]. Evaluating how family history contributes to increased disease risk is undoubtedly vital for establishing the effectiveness of family history information in the practice of preventive medicine and for developing public health strategies.

The first three chapters of my thesis present methods for comprehensively and systematically assessing phenotypes associated with family health history. In *Chapter 1*, in a study termed "Family History-Wide Association Study" (FamWAS), we investigate how specific environmental factors (e.g., smoking pollutant exposure) and clinical

phenotypes (body mass index, creatinine) contribute to family history and performed a comprehensive search for candidate phenotypes and environmental exposures associated with a family history of diabetes, asthma, and coronary heart disease from 457 clinical and environmental quantitative traits, including anthropometric and laboratory (e.g. urine and blood analysis) measures. In a similar methodology to Genome-Wide Association Studies (GWAS) where associations between genetic variants and disease phenotypes are evaluated across the entire genome, FamWAS evaluates traits associated with a family history across a broad scale of clinical and environmental quantitative traits and shines light on factors that have not previously been studied in family risk for disease. By conducting a systematic search across three different disease phenotypes, we allow for discovery of traits associated with multiple disease family history indicators, enhancing our understanding of disease similarity.

In *Chapter 2*, in a joint collaboration with Dr. Muin Khoury from the Centers for Disease Control and Prevention (CDC), we perform a population-based investigation assessing the combined influence of family history of CVD and diabetes on the prevalence of these diseases and on cardiovascular risk factors. In *Chapter 3*, in a study termed "XY-FamWAS", we perform 132 cross disease-family history associations for every disease X and family history of a different disease Y in order to gain potential insight into shared familial influences between 12 complex human diseases. We extend our study to deconvolve the genetic and environmental components of cross disease-family history associations, uncovering overlap of shared genetic architecture and environmental components by which unexpected diseases are related. Our studies in *Chapters 1, 2, and*

3 highlight the clinical utility and public health importance of family history information in assessing disease risk for early detection and potential intervention.

While observational studies have been instrumental in identifying risk factors in health and have paved the way to identifying appropriate interventions and therapeutics, determining biological relevance in the presence of confounding factors remains a challenge. Mendelian randomization, for which we present a protocol in *Chapter 4*, is a promising statistical method that can shed light on causal relationships by using genetic variants as instrumental variables (IV) for an exposure of interest.

In *Chapter 5*, we leverage Mendelian randomization to assess causality of associations identified in a data-driven observational study. Our two-pronged approach demonstrates a novel framework for studying the bidirectional link of diabetes and infectious diseases in large claims data and leverages genetic variation to examine putative causal relationships between infection and diabetes. In this study, we first conduct a data-driven systematic search, termed "Infection-Wide Association Study" (IWAS), for performing a comprehensive and systematic analysis for infectious disease in association with type 2 diabetes both before and after diagnosis from a nationally representative dataset of 44.9M members from large health insurance claims in the United States. Our approach allows for discovery of infectious diseases that may be possible risk factors of type 2 diabetes and/or complications, enhancing our understanding of the bidirectional link between type 2 diabetes and infectious diseases and prioritizing infectious diseases associated with type 2 diabetes warranted of further study. Second, we link infectious disease susceptibility

genotypes with type 2 diabetes using Mendelian Randomization. Our method advances our understanding of a genetic predisposition for infectious disease in increasing susceptibility for type 2 diabetes for infections identified from a broad spectrum of over 250 infectious diseases presented in type 2 diabetes, assessing the potential etiological role of infections in type 2 diabetes and providing robust evidence on the characterization of infection and type 2 diabetes that can help develop better clinical guidelines in type 2 diabetes management.

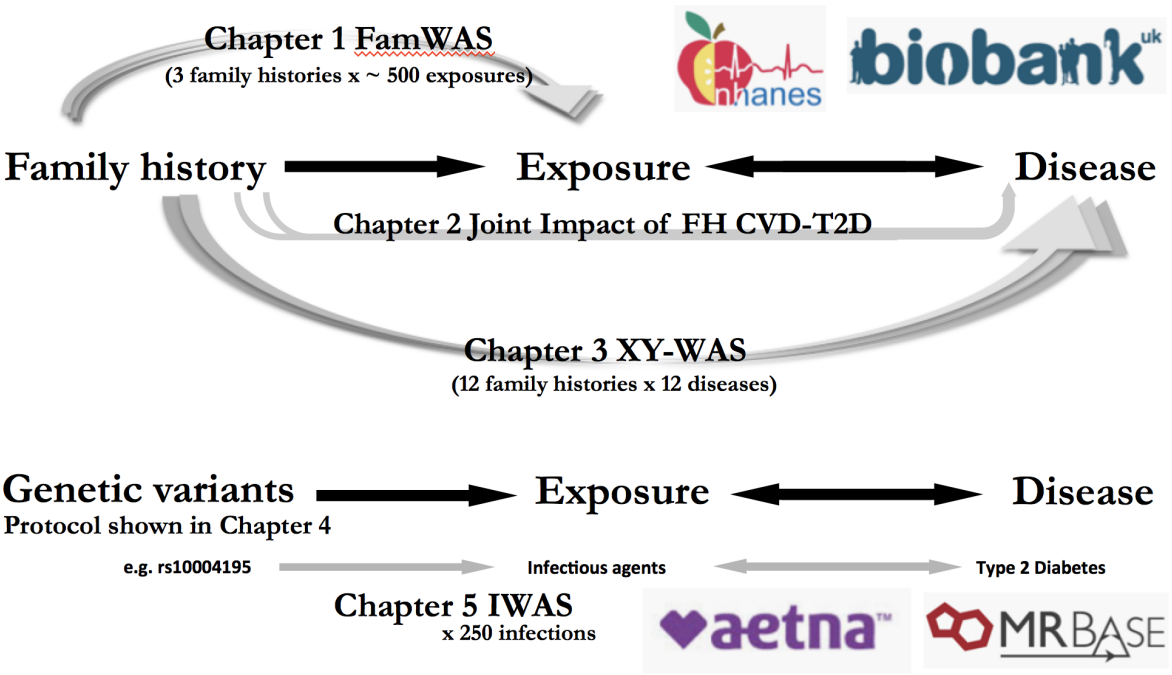


Figure 1.1. Schematic of specific aims and datasets.

The studies presented in this thesis suggest how harnessing family history information and the Mendelian randomization technique can be effective strategies for evaluating health risks in the practice of preventive medicine and public health.

1

Family History-Wide Association Study (FamWAS) for Type 2 Diabetes, Asthma, and Coronary Heart Disease

Family History–Wide Association Study to Identify Clinical and Environmental Risk Factors for Common Chronic Diseases [3]

Collaborators/co-authors: John PA Ioannidis, Muin J Khoury, Chirag J Patel
Manuscript published in *American Journal of Epidemiology* in August 2019

ABSTRACT

Family history is a strong risk factor for many common chronic diseases and summarizes shared environmental and genetic risk, but how this increased risk is mediated is unknown. We developed a "family history–wide association study" (FamWAS) to systematically and comprehensively test clinical and environmental quantitative traits (CEQTs) for their association with family history of disease. We implemented our method on 457 CEQTs for association with family history of diabetes, asthma, and

coronary heart disease (CHD) in 42,940 adults spanning 8 waves of the 1999–2014 US National Health and Nutrition Examination Survey. We conducted pooled analyses of the 8 survey waves and analyzed trait associations using survey-weighted logistic regression. We identified 172 (37.6% of total), 32 (7.0%), and 78 (17.1%) CEQTs associated with family history of diabetes, asthma, and CHD, respectively, in sub cohorts of individuals without the respective disease. Twenty associated CEQTs were shared across family history of diabetes, asthma, and CHD, far more than expected by chance. FamWAS can examine traits not previously studied in association with family history and uncover trait overlap, highlighting a putative shared mechanism by which family history influences disease risk.

Introduction

Family history is a strong risk factor for many common chronic diseases and summarizes shared environmental and genetic risk, but how this increased risk is mediated is unknown. We developed a "Family History-Wide Association Study" (FamWAS) to systematically and comprehensively test Clinical and Environmental Quantitative Traits (CEQTs) for their association with family history of disease. We implemented our method on 457 CEQTs for association with family history of diabetes, asthma, and coronary heart disease (CHD) in 42,940 adults spanning 8 waves of the 1999-2014 National Health and Nutrition Examination Survey (NHANES). We conducted pooled analyses of the 8 survey waves and analyzed trait associations using survey-weighted logistic regression. We identified 172 (37.6% of total), 32 (7.0%), and 78 (17.1%) CEQTs associated with family history of diabetes, asthma, and CHD, respectively, in sub-cohorts of individuals without the respective disease. 20 associated CEQTs were

shared across family history of diabetes, asthma, and CHD, far more than expected by chance. FamWAS can examine traits not previously studied in association with family history and uncover trait overlap, highlighting a putative shared mechanism by which family history influences disease risk.

Family history is a well-known risk factor for developing many common chronic diseases such as diabetes, asthma, and coronary heart disease (CHD) and reflects inherited genetic and shared environmental contribution in disease. Methods to delineate the mechanisms by which family history of disease influences inherited traits and environmental variables can be valuable in identifying how disease risk is conferred and distinguishing possible target areas amenable to intervention.

While previous efforts have studied the association between several specific candidate factors of disease and a family history, there may be as yet many undiscovered traits associated with a positive family history. We present here a "Family History-Wide Association Study" (FamWAS) to comprehensively identify Clinical and Environmental Quantitative Traits (referred herethereafter as 'CEQT') associated with family histories of chronic disease, focusing here on diabetes, asthma, and CHD. FamWAS extends previous studies that assess a few traits at a time with a single family history by systematically evaluating 457 CEQTs, including anthropometric and laboratory measurements as well as environmental pollutants, nutrients, and organic substances in association with three family histories in participants in the Continuous National Health and Nutrition Examination Survey (NHANES). Complementary to Genome-wide Association Studies (GWASs) or Phenome-wide Association Studies (PheWAS) methodologies[4,5],

FamWAS scans for traits associated with a family history in an unbiased approach on a broad scale while controlling for multiple testing, potentially uncovering novel associations that could enhance our understanding of the influence of family history. Some of the associated factors may be the ones that eventually mediate the increase in disease risk that family history is known to confer.

METHODS

NHANES cohort construction

We derived the cross-sectional study cohort from questionnaire and laboratory examinations of 8 independent waves of the 1999-2014 NHANES. Individuals selected to participate in NHANES were identified via a random sampling method and conducted questionnaires on health status, as well as clinical phenotypic measurements (e.g. body mass index) and laboratory tests (e.g. blood and urine analyses)[6].

For each respondent, we obtained demographic information, including age, sex, and race (collectively "covariates"), family history information and current disease status for diabetes, asthma, and CHD **(Figure 1.1, continued)**. These conditions were chosen because of the availability of questionnaire family history data in all 8 waves of the NHANES. Family history of each disease was ascertained from the Medical Conditions Questionnaire (MCQ) with an affirmative response to the question "were any of your biological that is, blood relatives including grandparents, parents, brothers, sisters ever told by a health professional that they had" for diabetes, asthma, and heart attack or

angina (referred herethereafter as “coronary heart disease [CHD]”). Current diabetes status was ascertained in two ways: (1) individuals diagnosed with diabetes were ascertained using an affirmative response to the question “have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes”, and (2) individuals with undiagnosed diabetes were ascertained from fasting glucose levels greater than 7.0 mmol/L (126 mg/dL) following at least a 6-hour fast or glycated hemoglobin greater than 6.5%, in accordance with the American Diabetes Association guidelines [7]. Current status for asthma was ascertained using an affirmative response to “has a doctor or other health professional ever told you that you have asthma”, or a forced expiratory volume in one second to forced vital capacity (FEV1/FVC) ratio less than 0.70, indicative of airflow obstruction[8]. We identified individuals diagnosed for the condition CHD by an affirmative response to diagnosis of congestive heart failure (CHF), CHD, angina, or heart attack. To ensure consistent reporting of family history and disease status, individuals with no information on family history or current disease status (a response of “Refused” or “Don’t know”) were removed from further analysis. Pregnant women and participants under 18 were also removed.

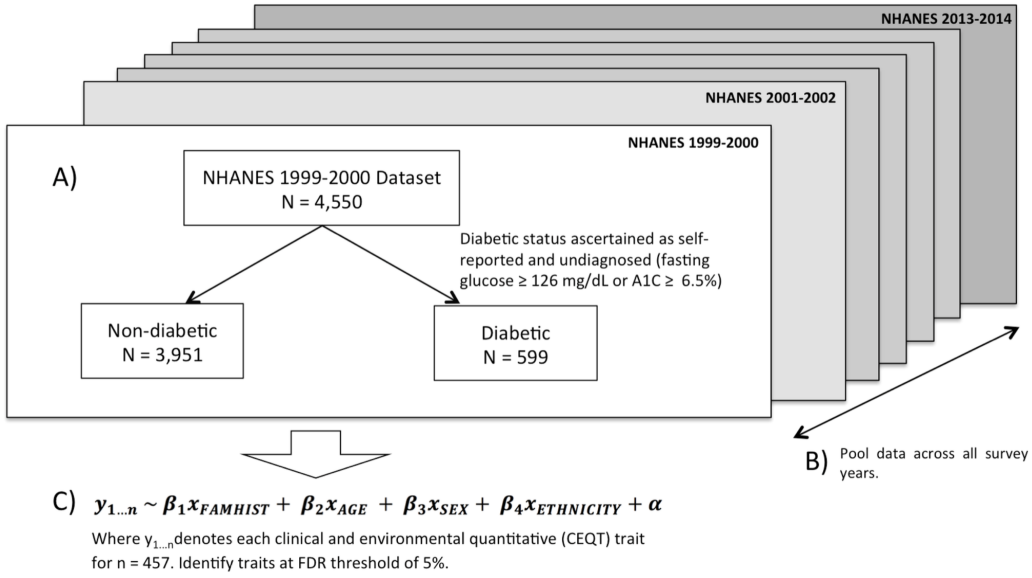


Figure 1.1. Method overview for pooled analysis. (A) We ascertained a diabetic disease status by a self-reported diagnosis or by a fasting glucose value 126 mg/dL) or glycated hemoglobin HbA1c value greater 6.5% in the laboratory testing panels for participants who did not self-report a diabetes diagnosis in 1999-2014 National Health and Nutrition Examination Survey (NHANES). (B) We pooled data across 8 independent waves of the NHANES spanning 1999-2014. (C) We systematically tested 457 clinical and environmental quantitative traits (CEQTs) for association with family history of diabetes, adjusting for age, sex, and race, in non-diabetic individuals using a survey-weighted logistic regression model. We identified traits that met an FDR threshold of 5%. We repeated this workflow [A-C] for family history of asthma and coronary heart disease, where disease status for asthma was ascertained by self-reported diagnosis or FEV1/FVC ratio less than 0.70 and coronary heart disease was ascertained by self-reported diagnosis of congestive heart failure, coronary heart disease, angina/angina pectoris, or heart attack.

Clinical and environmental quantitative traits selection

We collected a total of 457 CEQTs that represented a range of anthropometric, laboratory, and environmental attributes from the categories in **Table 1.1 (Continued)**. We removed non-continuous traits (e.g. languages spoken at home), and for measurements represented multiple times in different units (e.g. triglycerides measured in

mmol/L and mg/dL), we removed all but one measurement. To increase statistical power, traits with measurements present in less than 5% of the total population were also removed (e.g. osteoporosis-related measures). We investigated the summary statistics for each CEQT and plotted distributions using the raw values for each CEQT, and additionally scaled and log (base 10) transformations of each CEQT. We then fit transformations on a case-by-case basis appropriate for the distributions of each trait. CEQTs with approximate normal distributions were scaled and centered in analyses (mean-subtracted and divided by the standard deviation) to allow comparison of association sizes, which reflect a change per 1 standard deviation of the CEQT. CEQTs with right-skewed distributions were additionally log (base 10) transformed and scaled.

Table 1.1. Clinical and environmental categories for selection of 457 CEQTs

Clinical Traits	Category	No. of traits
	Bacterial infection	**9**
	Biochemistry	**42**
	Blood	**14**
	Blood pressure	**3**
	Body measures	**63**
	Cognitive functioning	**1**
	Cotinine	**2**
	Hormone	**13**
	Nutrients	**29**
	Physical fitness	**2**
	Phytoestrogens	**6**
	Spirometry	**10**
	Viral infection	**10**
	Total no. of clinical traits	**204**
Environmental traits	**Dialkyl**	**7**
	Dioxins	**7**
	Furans	**10**
	Heavy metals	**36**
	Hydrocarbons	**21**
	PCBs	**35**
	Perchlorate	**3**
	Perfluorochemicals	**12**
	Pesticides	**64**
	Phenols	**8**
	Phthalates	**15**
	Volatile compounds	**35**
	Total no. of environmental traits	**253**

Statistical analyses

To test for association between family history of disease and prevalent disease, we used survey-weighted logistic regression, adjusting for covariates. To account for the stratified and weighted design of NHANES, all regression models were fit using the *svyglm* function from the *survey* package in R[9].

For our main analyses, we tested each of the 457 CEQTs for association with family history of diabetes, asthma, and CHD in the pooled NHANES survey data using survey-weighted linear regression **(Figure 1.1B and C**). We constructed weights appropriate for combining 16 years of data that include 1999 through 2014. Notably, in the main analysis, we evaluated family-history-CEQT associations in individuals who were not diagnosed with the disease (removing also undiagnosed individuals with disease, as ascertained above). The reason is that we wanted to avoid the situation where disease may have affected some of these CEQTs. We wanted to use the FamWAS approach to reveal correlates of family history, some of which may then act also as risk factors for developing disease, rather than vice versa (disease being a risk factor for these correlates). Following the regression analyses, we used the false discovery rate method to correct for multiple testing [10]. We also evaluated for comparison family-history-CEQT associations on the entire cohort of individuals, regardless of participant disease status. We computed the number of identified traits shared between the three family histories and compared to the number of expected traits. The expectation was derived by taking the product of the three proportions of CEQTs identified for each disease, such that each event is assumed to be independent of the other two events, and calculating the product of that probability and the total number of CEQTs.

In addition to the pooled analysis, we conducted a meta-analysis of trait associations across survey years in order to ascertain heterogeneity of CEQTs across the different survey years due to potential year-by-year variation in disease prevalence. We tested each of the 457 CEQTs for association with family history of diabetes, asthma, and CHD in each wave of the NHANES survey using survey-weighted logistic regression, and then

meta-analyzed the associations for each trait and family history of disease across all survey years**.** All meta-analyses were conducted using the *metafor* package in R using a random effects meta-analysis [11]. We conducted a statistical test of heterogeneity to determine the variation among the association sizes observed for each trait across survey years.

RESULTS

Participant demographics

The initial cohort size was 82,091 participants. We excluded individuals who were under age 18 (34,735 individuals) or pregnant (another 1,182 individuals). To obtain the size for each disease cohort, we excluded participants with missing information on current disease status of diabetes, asthma, and CHD (32 adults for diabetes, 44 for asthma, and 3,916 for CHD), and family history (another 4401, 3190, and 1234 adults, respectively), which resulted in final cohort sizes for study of 41,741 eligible participants for diabetes, 42,940 for asthma, and 41,024 for CHD **(Table 1.2, Continued)**. Participants who responded "don't know" or "refused" are counted as missing information on family history and current disease status. The weight- and stratification-adjusted demographics and characteristics of each cohort are shown in **Table 1.2 (Continued)**.

Table 1.2. Demographic breakdown of diabetes, asthma, and CHD cohorts presented as weighted percentages, National Health and Nutrition Examination Survey, 1999-2014.

	Diabetes			Asthma			CHD		
Characteristic	All (41,741)	Has Diabetes[a] (6,324)	No Diabetes (35,417)	All (42,940)	Has Asthma[b] (6,886)	No Asthma (36,054)	All (41,024)	Has CHD[c] (3,527)	No CHD (37,497)
Age[d]	46.24	58.41	44.77	45.68	46.30	45.56	46.09	63.95	44.89
Female	51.54	49.02	51.81	51.23	54.01	50.76	51.55	43.29	52.08
Race									
White	69.18	61.15	70.16	68.87	73.44	67.90	69.20	77.70	68.60
Black	11.35	16.37	10.74	11.46	11.89	11.36	11.33	10.59	11.39
Mexican	7.95	8.98	7.82	8.08	4.33	8.86	7.97	4.00	8.24
Other Hispanic	5.45	6.18	5.36	5.48	4.92	5.90	5.43	3.18	5.58
Other	6.07	7.32	5.92	6.11	5.41	6.26	6.01	4.55	6.17
Positive family history	42.20	67.00	39.02	21.26	35.90	18.03	13.79	25.37	13.04

Abbreviations: CHD, coronary heart disease.

[a]Diabetic cases were classified as diagnosed (self-reported a diagnosis by a doctor or other health professional) or undiagnosed (fasting glucose value greater than 126 mg/dL or glycated hemoglobin HbA1c value greater than 6.5% in the laboratory testing panels for participants who did not self-report a diabetes diagnosis).

[b]Asthmatic cases were classified according to self-reported diagnosis or FEV1/FVC ratio < 0.70.

[c]Participants with coronary heart disease were classified according to self-reported diagnosis for congestive heart failure, coronary heart disease, angina/angina pectoris, or heart attack.

[d]Values are expressed as mean years.

Family history as a risk factor for diabetes, asthma, and CHD

Family history is a well known risk factor for diabetes, asthma, and CHD[12–14]; as expected, we observed a substantially increased risk of disease associated with a self-reported family history that was consistent across all survey years (**Figure 1.3, Continued**). The odds ratio (95% CI) for disease in association with a family history was 3.60 (3.28-3.96), 2.50 (2.33-2.67), and 2.68 (2.37-3.02) for diabetes, asthma, and CHD, respectively. The proportion of individuals with family history was 42.2%, 21.3%, and 13.8% in the diabetes, asthma, and CHD cohorts, respectively (**Table 1.2**).

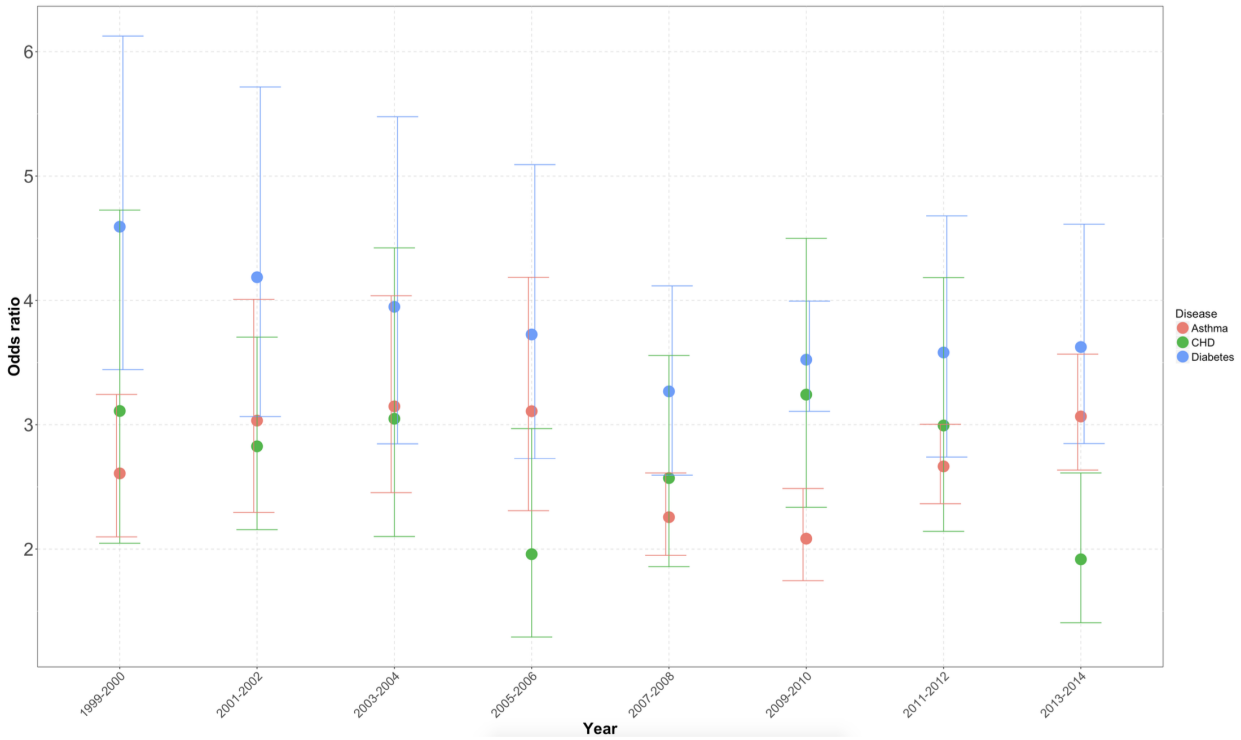


Figure 1.3. The trend for odds ratio for disease outcome (asthma in red, coronary heart disease in green, and diabetes in blue) in association with self-reported family history for each respective disease is shown in 8 waves of the NHANES (1999-2014).

Comprehensive search of CEQTs associated with a family history

We examined the summary statistics and distributions of each CEQT and applied case-by-case statistical transformations of each CEQT. **Figure 1.4 (Continued)** shows volcano plots illustrating the distribution of the CEQTs regression coefficients for family history of diabetes (**Figure 1.4A, Continued**), asthma (**Figure 1.4B, Continued**), and CHD (**Figure 1.4C, Continued**) in individuals without the respective disease. The traits at an FDR less than 5% are annotated on the plots. In the pooled analysis, of 457 tested CEQTs, 172 (37.6% of total), 32 (7.0%), and 78 (17.1%) CEQTs achieved an FDR of 5% for association with family history of diabetes, asthma, and CHD, respectively. The majority of the CEQTs exhibited little variation in study outcomes between NHANES

waves, with 38.3%, 40.6%, and 33.3% of the total number of traits at FDR of 5% having an I^2 estimate greater than 25% for diabetes, asthma, and CHD, respectively (**Figure 1.5, Continued**). The meta-analysis of CEQT-family history associations across survey years indicated relatively little heterogeneity (**Supplementary Figures 1.1-1.3**).

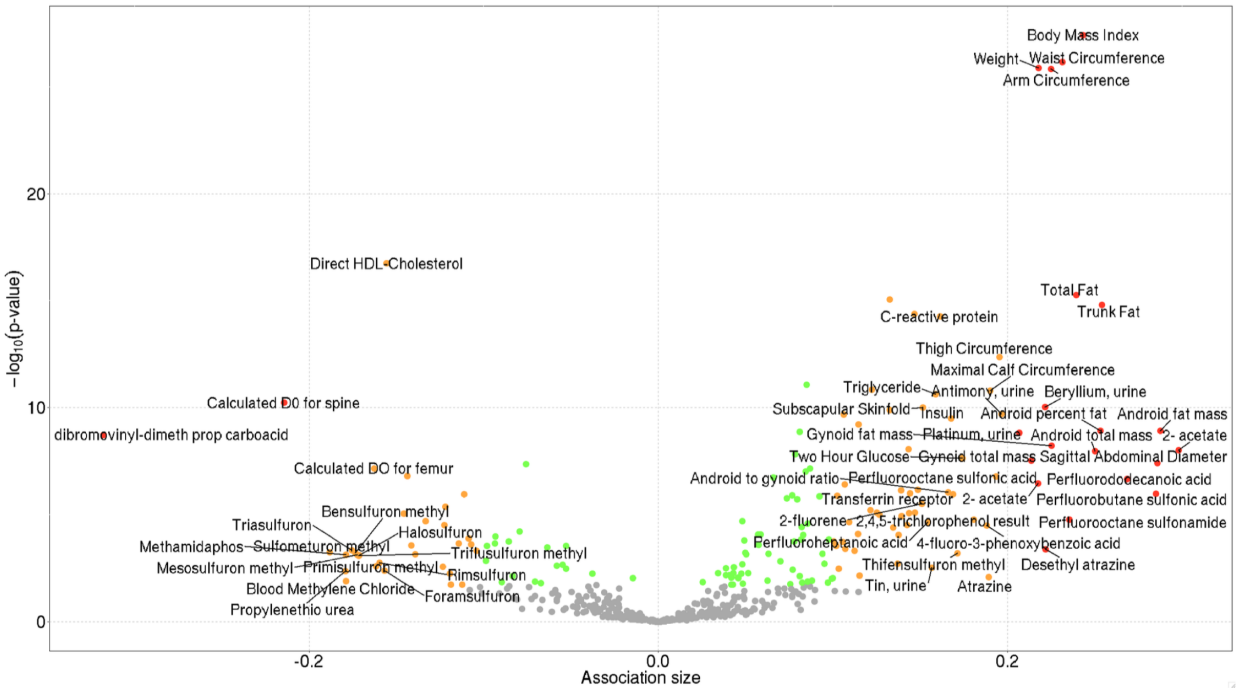


Figure 1.4A. Volcano plot results for pooled analysis in individuals without the respective disease. 457 CEQT association sizes versus -log10(p-value) for family history of diabetes (Figure 1.4A), asthma (Figure 1.4B), and coronary heart disease (Figure 1.4C) in individuals without the respective disease, adjusting for age, sex, and race, in 1999-2014 National Health and Nutrition Examination Survey (NHANES). Green, orange, and red points represent traits that met an FDR threshold of 5%. Orange and red points represent traits with an absolute value of association size greater than or equal to 0.10 and 0.20, respectively. All labeled points are traits that met an FDR of 5% and have an absolute value of association size greater than or equal to 0.15 in Figure 1.4A, 0.02 in 1.4B, and 0.08 in 1.4C.

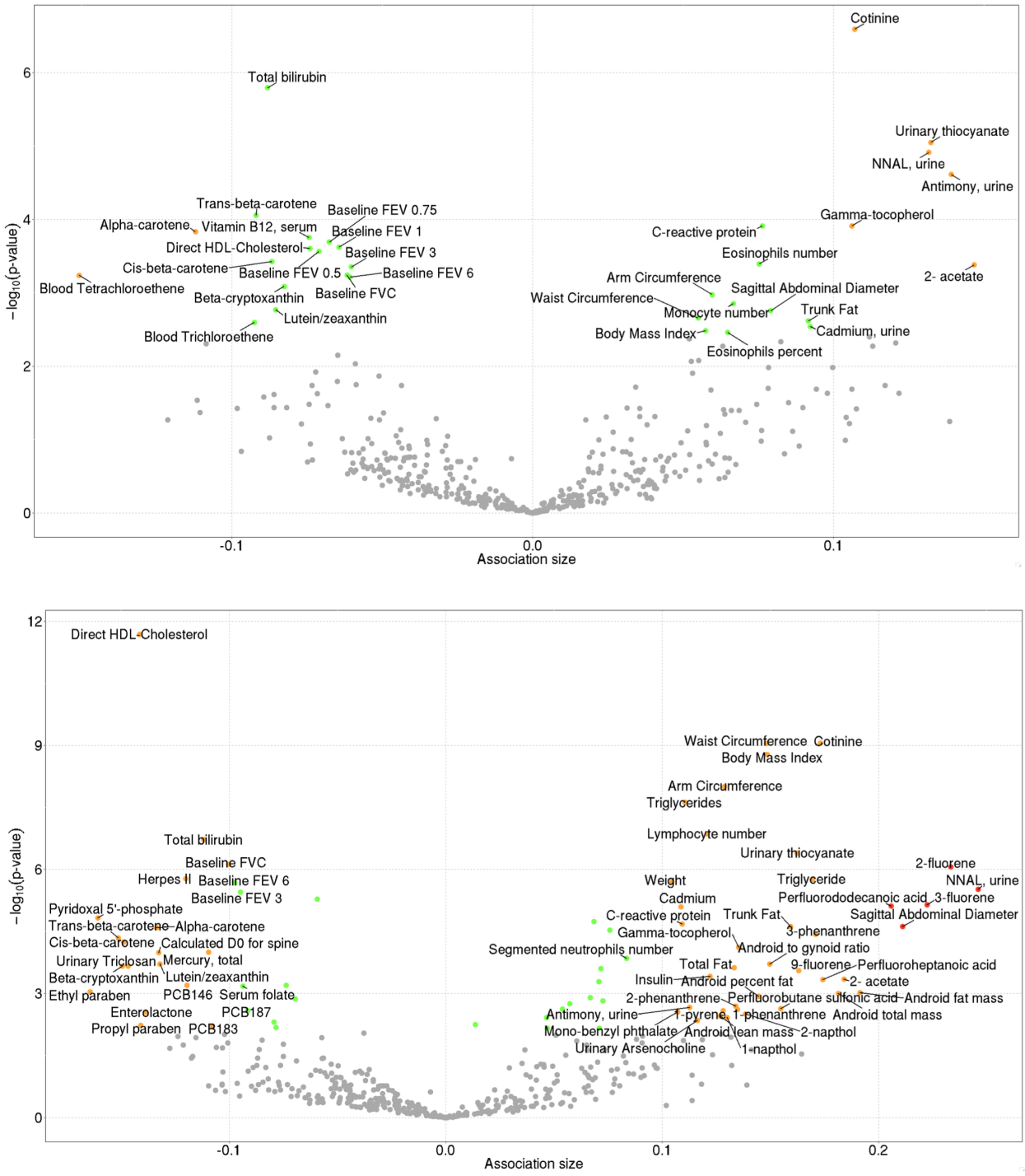


Figure 1.4B (top) and 1.4C (bottom). Volcano plot results for pooled analysis in individuals without the respective disease. 457 CEQT association sizes versus -log10(p-value) for family history of asthma (Figure 1.4B) and and coronary heart disease (Figure 1.4C). All labeled points are traits that met an FDR of 5% and have an absolute value of association size greater than or equal to 0.02 in 1.4B and 0.08 in 1.4C.

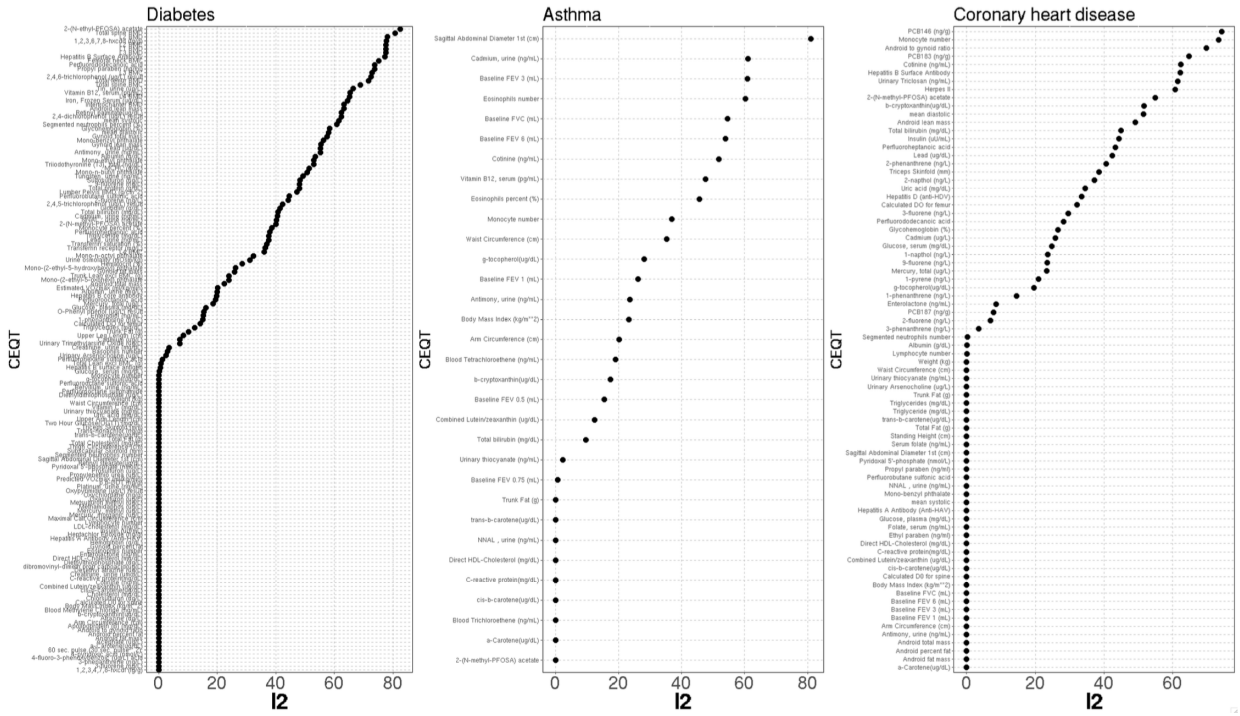


Figure 1.5. Heterogeneity estimates for CEQTs. I^2 estimate of associations between survey waves displayed for CEQTs that met an FDR threshold of 5% for diabetes (left), asthma (middle), and coronary heart disease (right) in individuals without the respective disease.

HDL-cholesterol and adiposity-related traits such as body mass index and waist circumference achieved the highest magnitude of association size for a family history of diabetes and CHD (**Figures 1.4A and 1.4C**). Cotinine (0.11 [0.09, 0.13]; *P* = 2.6e-7), urinary thiocyanate (association size 0.13 95%CI [0.11, 0.16]; *P* = 9.0e-6), and 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL) (0.13 [0.11, 0.16]; *P* = 1.2e-5), biomarkers of smoking, were identified in the asthma analyses, indicating that individuals not diagnosed with asthma or with airway obstruction but with a family history of asthma exhibited higher levels of tobacco smoke biomarkers (**Figure 1.4B**). Respiratory

measurements including baseline FEV from 0.5 seconds (-0.071 [-0.089, -0.053]; *P* = 2.7e-4) to 6 seconds (-0.061 [-0.077, -0.044]; *P* = 6.1e-4) and baseline FVC (-0.062 [-0.078, -0.045]; *P* = 5.7e-4) were also associated with a family history of asthma (**Figure 1.4B**).

We have identified an inverse association between pyridoxal 5'- phosphate (-0.144 [-0.166, -0.121]; *P* = 1.6e-7), the active form of vitamin B6, and a positive association between 2-fluorene (0.151 [0.120, 0.181]; *P* = 3.1e-6), a polycyclic aromatic hydrocarbon, with a family history of diabetes (in individuals without diabetes) (**Figure 1.4A**). We have also identified combined lutein/zeaxanthin associated with a family history of diabetes (-0.122 [-0.14, -0.099]; *P* = 4.1e-6) and CHD (-0.132 [-0.164, -0.100]; *P* = 1.9e-4) in individuals without the respective disease (**Figure 1.4A and 1.4C**). We found the volatile compounds blood tetrachloroethene (-0.151 [-0.192, -0.101]; *P* = 5.8e-4) and blood trichloroethene (-0.093 [-0.122, -0.063]; *P* = 2.5e-3) negatively associated with a family history of asthma in individuals without asthma (**Figure 1.4B**). We found cadmium (0.092 [0.062, 0.123]; *P* = 2.9e-3), a heavy metal, positively associated with a family history of asthma in individuals without asthma, as well as white blood cell count, measured by eosinophil number (0.075 [0.055, 0.096]; *P* = 4.0e-4), and monocyte number (0.067 [0.046, 0.087]; *P* = 1.4e-3), (**Figure 1.4B**). Body mass index, cotinine, and HDL-cholesterol were identified as the traits with the lowest FDR in association with a family history of diabetes, asthma, and CHD, respectively (**Figure 1.4**).

Shared and distinct family-history associated traits between a cohort of individuals without the respective disease and the entire cohort

We examined the overlap of family history-associated traits in individuals without the respective disease (including diagnosed and undiagnosed individuals) and the entire cohort (individuals with and without disease). **Supplementary Figure 1.4** shows volcano plots for CEQT associations in the entire cohort. A majority of the traits identified in the cohort of individuals without disease overlapped with the traits identified in the entire cohort, with 161 of 172 (93.6%) traits overlapping in the diabetes analyses, 30 of 32 (93.8%) in asthma, and 74 of 78 (94.9%) in CHD. Notably, we noticed many of the traits exhibited a strong positive linear relationship, and this was consistent among all three family histories, as well as among all 457 traits. We identified 46, 23, and 12 traits that demonstrated discordance of results for the cohort of individuals without the respective disease and the entire cohort analyses for diabetes, asthma, and CHD, respectively.

Shared traits among family histories of diabetes, asthma, and CHD

We examined traits shared between the family histories as well as traits that were associated with one family history (FDR of 5%) and not with the others (**Supplementary Figures 1.5-1.6**). **Figure 1.6 (Continued)** shows the 20 shared CEQTs associated with family histories of diabetes, asthma, and CHD. Of the 20 shared traits, 13 traits were not highly correlated (Pearson correlation coefficient (ρ) < 0.50) with each other, which is more than the expected 1.2 non-correlated traits shared across all 3 diseases. For all 20 shared traits, each trait exhibited either a positive association or a negative association consistent among all three family histories (**Figure 1.6, Continued**). Of the shared traits that exhibited a positive association, 5 were adiposity-related measures (e.g. arm circumference, trunk fat, sagittal abdominal diameter), indicating a shared relationship

between the three family histories and obesity. Smoking biomarkers (e.g. cotinine and urinary thiocyanate), vitamin-related compounds (e.g. γ-tocopherol), and liver-related compounds (C-reactive protein and bilirubin), were also shared in association with the three family histories. The association sizes were almost always larger for diabetes and CHD than for asthma.

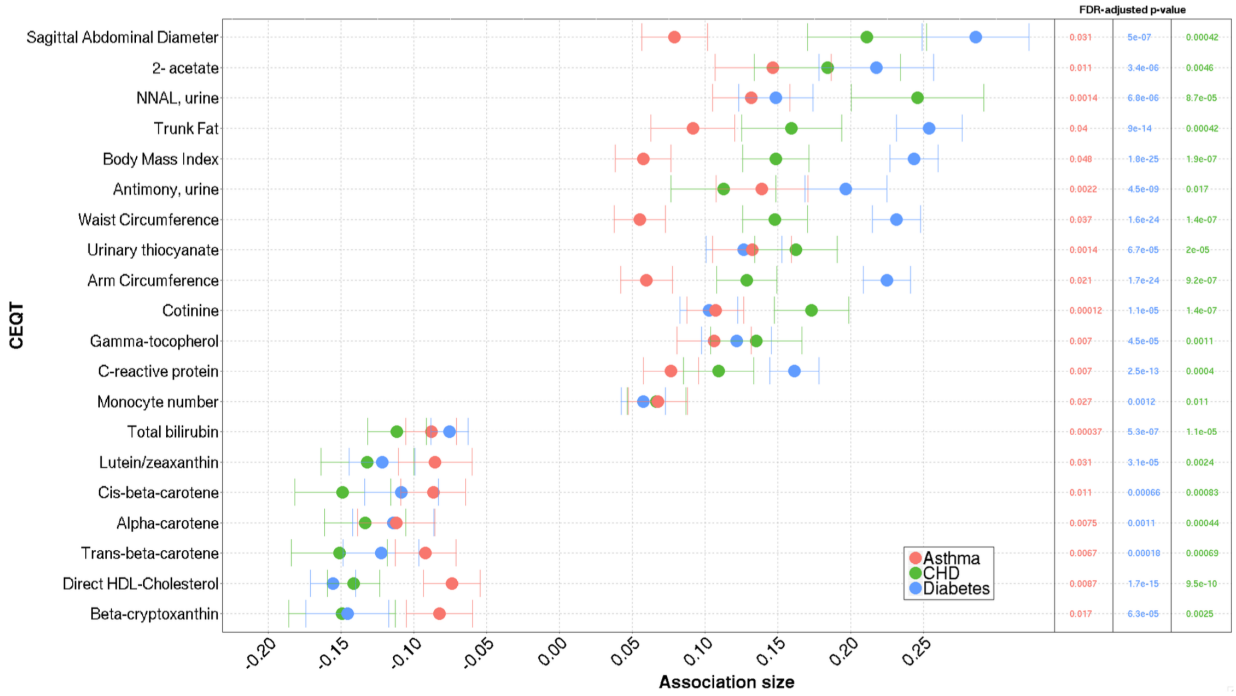


Figure 1.6. Shared CEQTs from pooled analysis. Shared CEQTs associated with family histories of diabetes (shown in blue), asthma (red), and coronary heart disease (green) in individuals without the respective disease in 1999-2014 National Health and Nutrition Examination Survey (NHANES). All CEQTs displayed achieved an FDR threshold of 5%. All models are adjusted for age, sex, and race. The FDR-adjusted p-value for each point is displayed to the right (in red for asthma, blue for diabetes, and green for coronary heart disease).

DISCUSSION

In this study, using a FamWAS approach, we comprehensively scanned 457 clinical and environmental quantitative traits for their association with family history of diabetes,

asthma, and CHD. By conducting a systematic search, we studied many phenotypes that have not been previously studied for their association to a family history, allowing for discovery of candidate phenotypes or environmental biomarkers. For example, we identified a novel association between decreased levels of pyridoxal 5'- phosphate, the biologically active form of vitamin B6, and family history of diabetes in individuals without diabetes, implicating that even in individuals with controlled levels of blood glucose and hemoglobin A1C, a positive family history can contribute to decreased levels of vitamin B6. We also identified a novel inverse association between lutein and zeaxanthin, carotenoids with antioxidant properties commonly found in egg yolks and green leafy vegetables, and a family history of diabetes and CHD in individuals without the respective disease. We further show that our method can lead to identification of traits associated among multiple disease family history indicators (e.g., family history associations shared between asthma, diabetes, and CHD), providing possible insight into the underlying biological similarities shared in the diseases [15].

While we have demonstrated the feasibility of a comprehensive search for traits associated with a family history of disease, we acknowledge that there are some limitations to our methodology. First, current disease diagnosis and family history of disease were ascertained using surveys, and self-reported measurements can be prone to measurement errors and recall biases. For example, participants may be underreporting family history of disease status, which may affect the estimates of CEQT-family history associations. Second, the NHANES family history questions pose several limitations. The blood relatives stated in the wording of the question groups together first-degree relatives (i.e., parent, sibling) as well as second-degree relatives (i.e., grandparents); however,

does not list all possible second-degree relatives, such as uncles, aunts, nephews, nieces, and grandchildren. This can result in a smaller population of individuals who reported a positive family history, and potentially an underestimation of CEQT-family history associations. Furthermore, the number of available family members and thus also the number of potentially affected family members will vary across participants and is not captured by the survey question.

Also, participants may have partial knowledge of their family history. Third, we excluded a number of participants due to missing information about family history and current disease status. However, even with the exclusion of participants, we obtained a large sample size of 42,940 eligible participants and the and the specificity of self-reported diabetes, asthma, myocardial infarction, and CHD ranged from 95% to 99%, while the sensitivity was 96% for diabetes, 91% for asthma, 90% for myocardial infarction, and 78% for CHD [16,17]. Furthermore, we estimated associations in individuals who were not, to the best of our knowledge, diagnosed with disease. For diabetes, we accounted for undiagnosed diabetes and mistaken reporting by marking participants without a reported diagnosis, but who fit ADA diagnostic criteria for diabetes, as individuals with diabetes for the purposes of all analyses. For asthma, we accounted for individuals with abnormal airway obstruction by marking participants with a FEV1/FVC ratio less than 0.70. For CHD, we marked participants who self-reported any of four cardiovascular events (CHF, CHD, angina, or heart attack) in order to ensure we marked individuals who had a condition with a symptomatic result of angina. 14.8% of the participants marked with a cardiovascular event had CHF and did not have any of the other three conditions. While CHF is not necessarily always due to CHD, we included these individuals because CHF

is often caused by coronary artery disease, heart attack, and other conditions that damage the heart muscle. In studies examining the accuracy of reported family history, a self-reported family history of diabetes when compared to physician-assessed diabetes status of close relatives had a sensitivity of 78.5% and specificity of 94.9%, a self-reported family history of CHD compared to status reported by parents had 85% sensitivity and 93% specificity, and a self-reported family history of asthma had 53% sensitivity and 99% specificity [18,19].

Another limitation includes a segment of the surveyed population that had missing data on family history. However, only 2.9-9.5% of individuals across the three cohorts met our selection criteria yet had missing data on family history. We speculate since most of the missing segment of the sample was of younger age, that the magnitude of the CEQT associations might be attenuated. Moreover, missingness was substantial for many traits, and in particular, for environmental variables, which can bias CEQT-family history associations and create larger uncertainty about these associations.

Family history may have several advantages over other analytic methods to find risk factors prior to disease onset because it reflects the complex interaction of shared genetic, environmental, lifestyle, and behavioral factors[2,13]. First, family history information is easy to capture and is commonly collected in population-based studies. A recent approach termed genome-wide association study by (GWAX) leveraged family history of disease information along with the genotypes of undiagnosed relatives to identify common genotypes in 12 common diseases, reconfirming known and identifying novel risk loci[20]. Their findings demonstrate the utility of family history to conduct

association mapping without direct case genotyping. Similarly, we leverage family history information in FamWAS to identify modifiable risk factors in addition to genetic risk factors prior to disease onset. A major strength reflected in FamWAS is in cases where a disease endpoint is unknown, family history information can be leveraged as a substitute in identifying modifiable risk factors shared in households. Our approach adds to the data-driven tools to identify phenotypes and exposures associated with family history. We show feasibility of our method by re-identifying previously known traits associated with a family history of the three prevalent chronic diseases.

Future directions include incorporating genotype information to partition DNA-transmitted genetic versus environmental variance in phenotype in family history to decompose the various components of risk influenced by familial disease, as we present in *Chapter 3* using data on up to 500,000 individuals from the UK Biobank. Further, we show feasibility in a cross-sectional dataset; FamWAS can be executed also on longitudinal cohorts including information on time of disease diagnosis for each participant in order to identify potential CEQTs that may mediate the association between family history and risk of disease. This may identify novel traits that can explain some of the remainder of the family history-associated disease risk, of which, for example, in diabetes, the known anthropometric and genetic risk factors currently explain a marginal ~20% of the association between family history and diabetes risk[21,22].Comparison of FamWAS results also across multiple datasets and cohorts with different settings and background could also further help understand the consistency of these associations.

2
Joint Impact of Family History of Diabetes and Cardiovascular Disease

The Joint Impact of Family History of Diabetes and Cardiovascular Disease among US adults: A Population-Based Study of 2007-2018 NHANES

Collaborators/Co-authors: Quanhe Yang, Ramal Moonesinghe, Gloria Beckles, Muin J Khoury, Chirag J Patel

ABSTRACT

Background. Family history is an established risk factor for both cardiovascular disease (CVD) and diabetes; however, to our knowledge, no study has presented population-based prevalence estimates of family histories of both CVD and diabetes and their joint impact on population prevalence of diabetes, heart disease and their risk factors.

Methods. We analyzed data from 29,440 participants aged 20 and over from six 2-year cycles of the National Health and Nutrition Examination Survey (NHANES 2007-2018) and assessed self-reported first-degree family history of diabetes and CVD (premature heart disease before age 50) as well as meeting criteria and/or having risk factors for CVD and diabetes. We performed survey-adjusted logistic regression to examine the association between family history of CVD and/or diabetes and a diagnosis of CVD and/or diabetes and CVD/diabetes risk factors.

Results. 44.4% of the US adult population have a family history of CVD and/or diabetes. Overall, 5.5%, 31.6%, and 7.3% of US adults have a family history of CVD without family history of diabetes, family history of diabetes without family history of CVD, and family histories of both conditions, respectively. Compared to those with no family history, having both diabetes and CVD family histories was associated with a prevalence ratio of 2.9, 2.6, and 5.3 for CVD, diabetes, and both diseases concurrently. Participants with both family histories of CVD and diabetes are diagnosed with diabetes 6.6 years earlier than individuals without family history of either disease or are at 6.5 greater odds for having both diseases in their lifetime.

Conclusion. A large proportion of individuals in the US have family history of both or any family history of CVD and diabetes that is comparable in risk to common cardiometabolic risk factors. This wide presence of high-risk family history and its simplicity of ascertainment suggests that clinical and public health efforts should harness family history information across CVD and diabetes to improve population efforts in the early detection and prevention of these common chronic diseases.

INTRODUCTION

We showed in *Chapter 1* that family history contributes to a slew of environmental and phenotypic risk factors for disease jointly. But how does combined family history play a role in multiple diseases? We explore this phenomenon in this and the following chapter, documenting new methods to exploit how a humble and easy-to-ascertain variable, family history, plays a role in diverse and multiple disease outcomes simultaneously, laying groundwork to study the complex etiology between genetics, shared environment, phenotype, and diseases.

Type 2 diabetes is a major risk factor for cardiovascular disease[23,24]. Further, the genetic and environmental antecedents shared by type 2 diabetes and cardiovascular disease (CVD) have led to the hypothesis that both arise from a "common soil"[25]. Family history, which reflects genetic susceptibility and shared environmental, behavioral, and lifestyle factors, can be used to investigate the cardiovascular disease continuum[26]. Previous studies have shown a family history of diabetes is associated with endothelial dysfunction and increased cardiovascular risk[27], and conversely, a higher familial risk of coronary heart disease is associated with incident type 2 diabetes among individuals with a positive family history for diabetes, while the association remains weak among individuals without a family history of diabetes[28]. A family history of premature heart or vascular disease is a known risk factor for cardiovascular events [29]. We have previously found that a family history of diabetes and coronary heart disease is associated with a broad array of risk factors[3,30]. However, the

combined role of family history of both CVD and diabetes on disease risk and cardiovascular health indicators is unclear. To our knowledge, no population-based investigation has assessed the combined influence of family history of premature CVD (under age 50) and diabetes on the prevalence of CVD and diabetes and their risk factors.

In this study, we analyzed a representative survey of the US population, National Health and Nutrition Examination Survey (NHANES) 2007-2018, to characterize the prevalence of family history of premature CVD (referred to hereafter as "family history of CVD") and diabetes. We further examined the association between family history of CVD and diabetes and prevalence of diagnosed CVD and diabetes, and CVD/diabetes risk factors based on criteria from the American Heart Association[31].

METHODS

Data source

The NHANES is conducted by the Centers for Disease Control and Prevention's National Center for Health Statistics (CDC/NCHS) and consists of cross-sectional studies designed to assess the health and nutritional status of noninstitutionalized, civilian residents from the US population with data released every other year in 2-year cycles[32]. The NHANES uses a questionnaire on medical conditions that includes information on demographic characteristics and family history of disease. In addition, physical examination and laboratory testing are administered by medical personnel in mobile examination center (MEC). The survey uses a complex sampling design to attain a sample that is representative of the non-institutionalized civilian population of the United States. We analyzed data for 29,440 males and non-pregnant females aged 20 and over

from six 2-year cycles of the 2007-2018 NHANES with available information on current disease status, first-degree family history of diabetes and coronary heart disease, and cardiovascular health risk factor measurements (**Figure 2.1**).

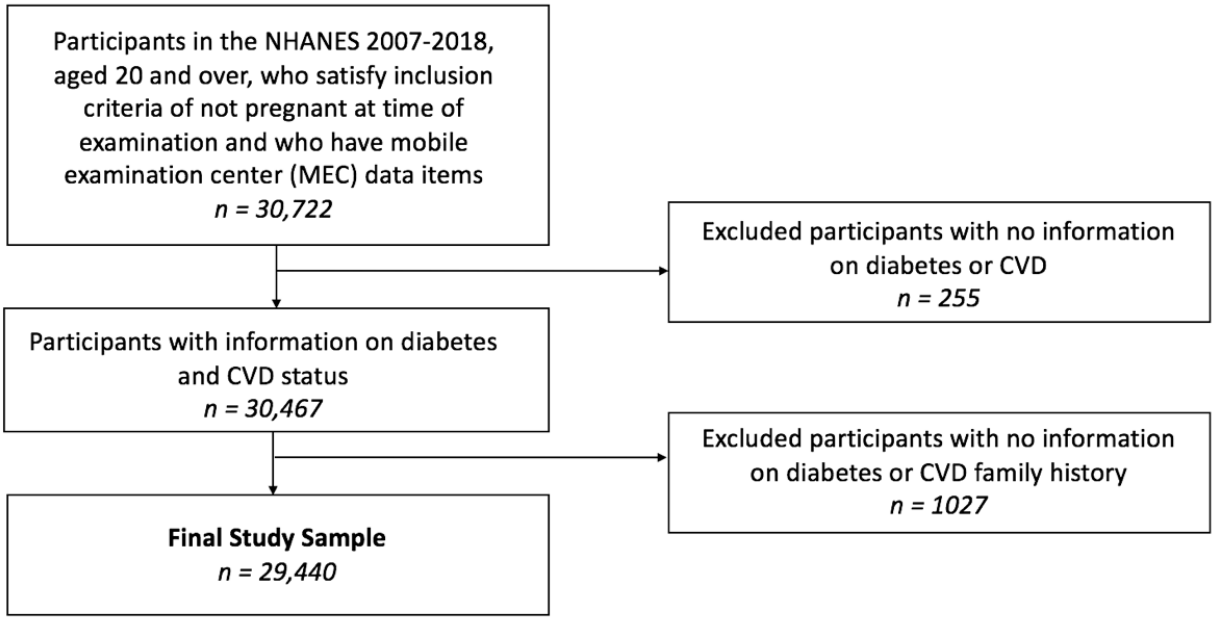


Figure 2.1. Flowchart for inclusion and exclusion criteria for study sample derivation, NHANES 2007-2018

Identification of participants with diabetes and CVD

We identified survey participants with diabetes using: (a) self-reported diagnosis of diabetes or (b) undiagnosed diabetes, identified in participants who do not self-report a diagnosis of diabetes but who meet a fasting glucose or hemoglobin A1C concentration level in accordance with the guidelines of the American Diabetes Association (ADA)[33]. Specifically, participants with diabetes met at least one of three criteria: (1) self-reported a physician diagnosis by an affirmative response to the question "have you ever been told by a doctor or health professional that you have diabetes or sugar

diabetes"; (2) a negative response to the question and fasting glucose level 7.00 mmol/L (126 mg/dL) or greater following at least an 8-hour fast, OR hemoglobin A1C concentration 6.5% or greater[33]. Since laboratory equipment for measuring glucose levels changed during the NHANES 2007-2018 period, we applied glucose regression equations as advised by the NCHS for consistency among estimates[34]. We identified participants diagnosed with CVD by an affirmative response to questions about the diagnosis of coronary heart disease, angina, heart attack, or stroke. Participants who reported "refused" or "don't know" to the current disease status questions, or with missing glucose laboratory measures were removed from analyses (n=255, 0.83%).

Identification of participants with family history of heart disease or diabetes

We identified individuals who reported a positive family history as those who had a first-degree affected relative (parent and/or sibling). NHANES ascertained family history of diabetes with an affirmative response to "Including living and deceased, were any of your close biological, that is, blood relatives including father, mother, sisters, or brothers, ever told by a health professional that they had diabetes" in the Medical Conditions Questionnaire. NHANES ascertained family history of CVD by an affirmative response to the question "Including living and deceased, were any of your close biological, that is, blood relatives including father, mother, sisters, or brothers, ever told by a health professional that they had a heart attack or angina before the age of 50." Participants who lacked knowledge or refused to respond to either question were removed from further analyses (n=1027, 3.34%). We classified participants according to four family history categories: (1) no family history of diabetes or CVD, (2) family history of CVD and no

family history of diabetes, (3) family history of diabetes and no family history of CVD, and (4) family history of both diabetes and CVD.

Demographic, behavioral, and clinical risk factors for CVD and diabetes

We estimated associations between family history of CVD and diabetes with risk factors, including demographic characteristics (age group [20-39, 40-59, 60+], sex, and race and Hispanic ethnicity [non-Hispanic white, Mexican-American, non-Hispanic black, and other Hispanic]), as well as measures of socioeconomic status, including poverty-income-ratio, computed as the ratio of family income to poverty threshold[35] and educational status of the participants (less than high school completion, or high school completion or greater).

We additionally estimated associations between family history and CVD/diabetes risk factors based on criteria from the American Heart Association, including body mass index (BMI) (<25, 25-29.9, 30 kg/m^2), smoking status, physical activity, total cholesterol, and blood pressure[31]. We defined participants as current smokers if the participant responded to the question "Do you now smoke cigarettes?" with "every day" or "some days" and as a non-smoker if the participant responded to the question with "Not at all." We defined participants as physically active if the participant reported at least 150 minutes a week of moderate-intensity or 75 minutes a week of vigorous intensity aerobic physical activity[36].

We also estimated associations between family history and high-density lipoprotein (HDL) and low-density lipoprotein (LDL) levels. In accordance with classifications set

by the National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III)[37], we categorized participants with a total cholesterol of <200, 200-239, and 240 as low/desirable, borderline high, and high, LDL cholesterol levels of <100, 100-159, 160 as low/desirable, borderline high, and high, and HDL cholesterol levels of <40, 40-60, 60 as low, borderline high, and high/desirable, respectively. We classified participants as having a normal blood pressure by a systolic blood pressure and a diastolic blood pressure , and as having hypertension by a systolic blood pressure and/or a diastolic blood pressure , or if they have controlled hypertension, defined as having a normal blood pressure but self-reporting that a doctor or health professional has diagnosed them with a high blood pressure[38]. We additionally accounted for C-reactive protein (CRP), a marker of systemic inflammation known to be associated with glycemic control and increased risk of diabetes, and identified participants with normal serum CRP level (less than 3 mg/L) and high CRP levels (more than 3 mg/L)[39–41].

Statistical analyses

We calculated the crude prevalence of family history of CVD, diabetes, and both according to various demographic characteristics and risk factors and estimated the total number of participants in each category.

We used logistic regression to quantify the associations between family history and prevalent disease, adjusting by age, sex, and race and ethnic origin. We estimated prevalence ratio (PR) for CVD and diabetes, which is inclusive of undiagnosed diabetes

and self-reported diagnosis of diabetes, given a family history, where findings are interpreted as the proportion of people with diagnosed CVD and diabetes given a family history over the proportion of individuals with no family history of CVD and diabetes. One model was constructed for each of four outcomes: prevalence of diabetes, CVD, both, and either/or. We additionally evaluated the association of family history with demographic characteristics and risk factors in participants who were not diagnosed with diabetes or CVD and who did not meet a fasting glucose or hemoglobin A1C level indicative of diabetes. In accordance with the NHANES analytic and reporting guidelines, all analyses accounted for the complex sample design by using design variables, and adjusting for MEC exam sample weights corresponding to pooling six 2-year survey cycles of the continuous 2007-2018 NHANES. To correct for multiple hypothesis testing for regression analyses, we calculated the false discovery rate (FDR) and denoted significance by an FDR threshold of 5%[10]. We conducted all statistical analyses using R software[42].

RESULTS

Prevalence of family history of CVD and diabetes

Our study included 30,722 participants who met the inclusion criteria of having MEC data items and not being pregnant at the time of examination. Of this study cohort, 255 (0.83%) participants were excluded due to no reported information on current diabetes or CVD status and a further 1027 (3.34%) participants were excluded due to no reported information on diabetes and CVD family history. The examination response rate for adult participants in these survey cycles was: 70.6, 72.2, 64.5, 63.7, 58.1, and 45.3[43]. The

prevalence of self-reported first-degree family history of CVD and no family history of diabetes, diabetes and no family history of CVD, and family histories of both CVD and diabetes was 5.5% (95% CI, 5.1-5.9%), 31.6% (95% CI, 30.8-32.5%), and 7.3% (95% CI, 6.8-7.8%), respectively **(Table 2.1, Continued)**. Trends in prevalence of family history during the period of 2007-2008 to 2017-2018 exhibited relative stability. Across all three positive family history categories, the prevalence of family history was higher in females than in males, and the prevalence of a family history of diabetes only or both CVD and diabetes was higher in populations with BMI greater than 30.

Table 2.1. Crude prevalence of reported family history of CVD and diabetes by selected demographic characteristics, NHANES 2007-2018.

Reported family history of:	Neither	CVD only[1]	Diabetes only[2]	Both
	Prevalence (%)	Prevalence (%)	Prevalence (%)	Prevalence (%)
Overall population	55.60 (54.66-56.54)	5.48 (5.09-5.87)	31.64 (30.83-32.47)	7.28 (6.77-7.79)
Age (years)				
20-39	61.61 (60.34-62.88)	3.80 (3.34-4.25)	29.02 (27.84-30.19)	5.58 (4.96-6.19)
40-59	51.98 (50.55-53.41)	5.61 (4.96-6.25)	33.81 (32.50-35.11)	8.61 (7.72-9.50)
60+	52.50 (50.80-54.21)	7.61 (6.71-8.51)	32.17 (30.89-33.46)	7.71 (6.83-8.60)
Sex				
Males	59.03 (57.89-60.15)	5.05 (4.60-5.51)	29.94 (28.94-30.93)	5.99 (5.32-6.66)
Females	52.97 (51.75-54.19)	5.80 (5.24-6.37)	32.96 (31.90-34.02)	8.27 (7.62-8.92)
Race/Ethnicity				
Non-Hispanic White	57.42 (56.27-58.57)	6.55 (5.97-7.13)	28.56 (27.57-29.54)	7.47 (6.79-8.16)
Mexican-American	50.67 (48.72-52.61)	2.76 (2.23-3.27)	40.43 (38.47-42.39)	6.14 (5.48-6.81)
Other Hispanic	59.17 (56.88-61.47)	3.32 (2.71-3.92)	31.15 (29.27-33.02)	6.36 (5.40-7.33)
Non-Hispanic Black	46.99 (45.30-48.70)	3.40 (2.79-4.00)	41.54 (40.10-42.98)	8.07 (7.34-8.79)
BMI (kg/m2)				
<25	63.67 (62.09-65.24)	5.49 (4.82-6.17)	25.78 (24.36-27.21)	5.06 (4.49-5.63)
25-29.9	57.96 (56.70-59.21)	5.82 (5.21-6.43)	30.15 (28.98-31.32)	6.07 (5.39-6.74)
30+	46.56 (45.13-47.98)	5.17 (4.64-5.68)	37.99 (36.64-39.33)	10.29 (9.45-11.13)
Poverty-Income-ratio				
< 1	52.29 (50.59-53.98)	5.10 (4.38-5.82)	33.33 (31.77-34.90)	9.28 (8.21-10.36)
≥ 1	56.50 (55.53-57.48)	5.58 (5.15-6.01)	31.19 (30.31-32.06)	6.73 (6.22-7.24)
Education				
<high school	51.93 (50.18-53.68)	6.13 (5.18-7.09)	32.59 (31.08-34.09)	9.35 (8.26-10.43)
≥ high school	56.31 (55.35-57.28)	5.35 (4.94-5.76)	31.46 (30.60-32.32)	6.87 (6.40-7.35)

[1]A reported family history of "CVD only" denotes a family history of CVD and no family history of diabetes.
[2]A reported family history of "diabetes only" denotes a family history of diabetes and no family history of CVD.

Prevalence of CVD and diabetes, by family history status

Further, participants with a family history of either CVD or diabetes were more likely to have the other condition. Specifically, there was a 2.36-fold (95% CI, 2.12-2.63) increase of having a family history of CVD among individuals with a family history of diabetes, adjusted by age, sex, and race/ethnicity (data not shown). Participants with a family history of CVD, diabetes or both had increased prevalence of having CVD, diabetes, both

conditions, or either condition compared to participants without a family history of either condition (**Table 2.2, Continued**). Compared to participants without a family history of CVD or diabetes, participants with a family history of CVD had a PR of 2.23 (95% CI, 1.90-2.61), 1.12 (95% CI, 0.91-1.37), 1.94 (95% CI, 1.33-2.82), and 1.53(95%CI, 1.35-1.74) for having CVD, diabetes, both, and either/or diabetes/CVD (**Table 2.2, Continued**). The PRs for participants with a family history of diabetes and not CVD were 1.38 (95% CI, 1.25-1.52), 2.35 (95% CI, 2.14-2.59), 2.76 (95% CI, 2.29-3.32), and 1.75 (95%CI, 1.62-1.90) for having CVD, diabetes, both CVD and diabetes, and either/or CVD and diabetes (**Table 2.2, Continued**). We found higher PRs of 2.85 (95% CI, 2.44-3.34), 2.60 (95% CI, 2.29-2.94), 5.27 (95% CI, 4.24-6.54), and 2.11 (1.90-2.35) for having CVD, diabetes, both, and either/or respectively, for participants with family histories of both CVD and diabetes.

Table 2.2. Prevalence (%) of CVD and diabetes among those in each family history category and prevalence ratios (95% CI) according to family history, adjusted by age, sex, race/ethnicity, BMI, education, and poverty-income-ratio, NHANES 2007-2018.

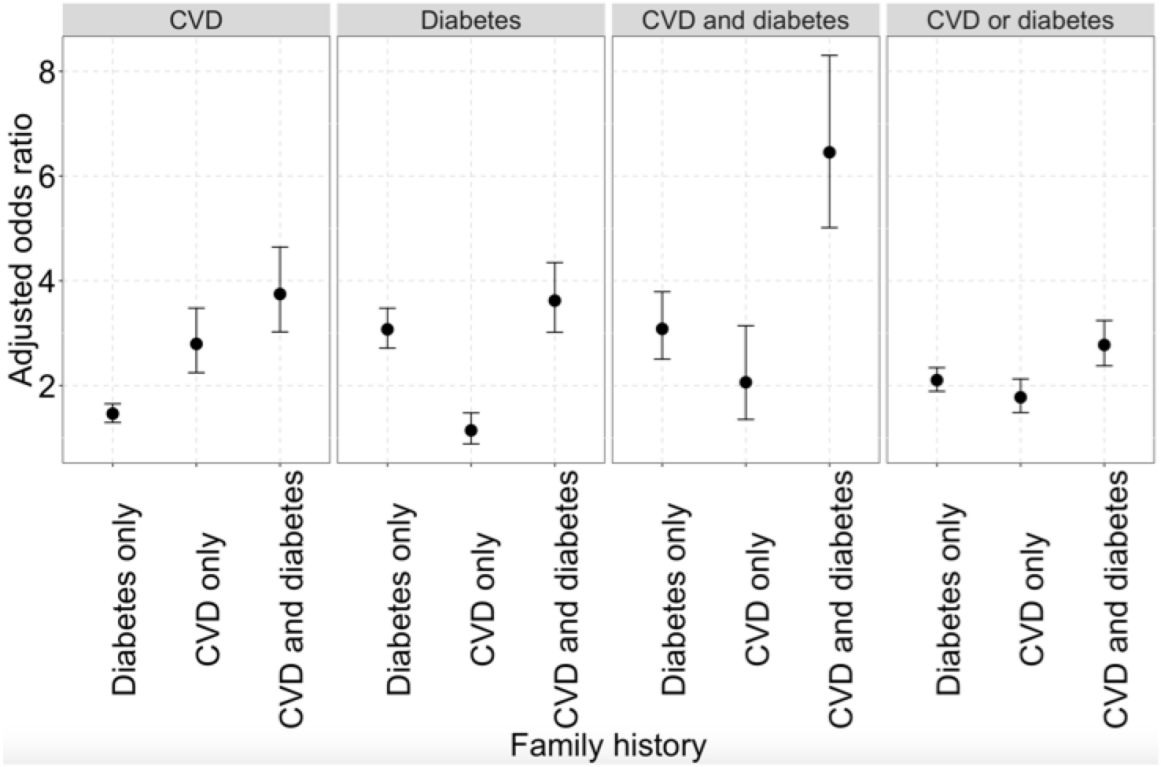
	No family history of CVD or diabetes	CVD family history only	Diabetes family history only	CVD and diabetes family history
CVD diagnosis				
Prevalence (%)	5.37 (4.95-5.79)	15.43 (12.83-18.02)	7.58 (6.96-8.19)	16.04 (13.92-18.16)
Prevalence ratio	Reference	2.23 (1.90-2.61)	1.38 (1.25-1.52)	2.85 (2.44-3.34)
Diabetes diagnosis				
Prevalence (%)	7.11 (6.49-7.73)	9.89 (8.03-11.56)	20.00 (18.85-21.16)	24.26 (21.79-26.73)
Prevalence ratio	Reference	1.12 (0.91-1.37)	2.35 (2.14-2.59)	2.60 (2.29-2.94)
CVD and diabetes diagnosis				
Prevalence (%)	1.26 (1.06-1.45)	3.17 (1.94-4.45)	3.74 (3.28-4.21)	7.36 (6.02-8.70)
Prevalence ratio	Reference	1.94 (1.33-2.82)	2.76 (2.29-3.32)	5.27 (4.24-6.54)
CVD or diabetes diagnosis				
Prevalence (%)	9.97 (9.25-10.68)	18.89 (16.41-21.36)	20.09 (18.96-21.22)	25.58 (23.44-27.71)
Prevalence ratio	Reference	1.53 (1.35-1.74)	1.75 (1.62-1.90)	2.11 (1.90-2.35)

† : Reference group is cohort of participants with no diagnosis of CVD and diabetes.
[1]Diabetes here is inclusive of diagnosed (self-reported a diagnosis by a doctor or other health professional) and undiagnosed (fasting glucose value greater than 126 mg/dL or glycated hemoglobin value greater than 6.5% in the laboratory testing panels for participants who did not self-report a diabetes diagnosis)

Risk factors for CVD and diabetes, by family history status

Supplementary Table 2.1 presents the adjusted odds ratios (aOR) for CVD and diabetes risk factors, including family history, age, sex, race and ethnic origin, BMI, income-to-poverty-ratio, and education. The left three columns present associations between risk factors with a diagnosis of CVD by family history status and the right three columns present the association between risk factors and having diabetes by family history status. The aOR for having a diagnosis of CVD, given a family history of CVD and not diabetes was 3.04 (95% CI, 2.65-3.50), and 3.37 (95% CI, 2.74-4.14) given a family history of both CVD and diabetes. The aOR of having CVD or diabetes given a family history of both CVD and diabetes was higher among males than females (1.79; 95% CI, 1.54- 2.09

for a diagnosis of CVD, and 1.52; 95%CI 1.33-1.73 for a diagnosis of diabetes) (Supplementary Table 2.1). **Figure 2.2**, rather than present aOR for risk factors, shows aORs for having CVD, diabetes, and both CVD and diabetes by family history status controlling for all risk factors shown in Supplementary Table 2.1. The aOR for having both CVD and diabetes was 3.08 (95% CI, 2.50-3.79), 2.06 (95% CI, 1.35-3.14), and 6.45 (95% CI, 5.01-8.30) in association with a family history of diabetes only, CVD only, and both CVD and diabetes, respectively (**Figure 2.2**).



† : Reference group is cohort of participants with no family history of CVD or diabetes.
[1]Diabetes here is inclusive of diagnosed (self-reported a diagnosis by a doctor or other health professional) and undiagnosed (fasting glucose value greater than 126 mg/dL or glycated hemoglobin value greater than 6.5% in the laboratory testing panels for participants who did not self-report a diabetes diagnosis)

Figure 2.2. Estimates of adjusted odds ratios (95% CI) for CVD, diabetes, and both CVD and diabetes according to family history of CVD, diabetes, both CVD and diabetes, and either CVD or diabetes, adjusted by age, sex, race/ethnicity, BMI, poverty-income-ratio, and education, NHANES 2007-2018.

Family-history associated prevalence and prevalence ratios for CVD and diabetes differ by risk factors, but remain similar across certain desirable and high categories for LDL-c, total cholesterol, and blood pressure for certain family history groups

The prevalence and PR for CVD and diabetes given the family history categories differed by levels of established risk factors for CVD and diabetes **(Figure 2.3, Figure 2.4, continued)**. Surprisingly, we found that the prevalence and PRs for CVD according to a family history of CVD was similar across low/desirable, borderline high, and high LDL cholesterol levels **(Figure 2.3a; Figure 2.4a, continued)**. Further, we also found that the PR for CVD according to a family history of diabetes was also similar across low/desirable, borderline high, and high total cholesterol levels **(Figure 2.4a, continued)**. The PR for CVD according to family histories of both CVD and diabetes remained similar across hypertensive and normal blood pressure levels **(Figure 2.4a, continued)**. For family histories of both CVD and diabetes, the prevalence for CVD for all three BMI categories remained similar **(Figure 2.3a, continued)**, while the same pattern was not noticed for the prevalence of diabetes (**Figure 2.3b, continued**). Further, for smoking status, for family histories of both CVD and diabetes, we identified that the prevalence of diabetes was similar for non-smokers and smokers (**Figure 2.3b, continued**), while the prevalence of CVD was higher among smokers than non-smokers (**Figure 2.3b, continued**).

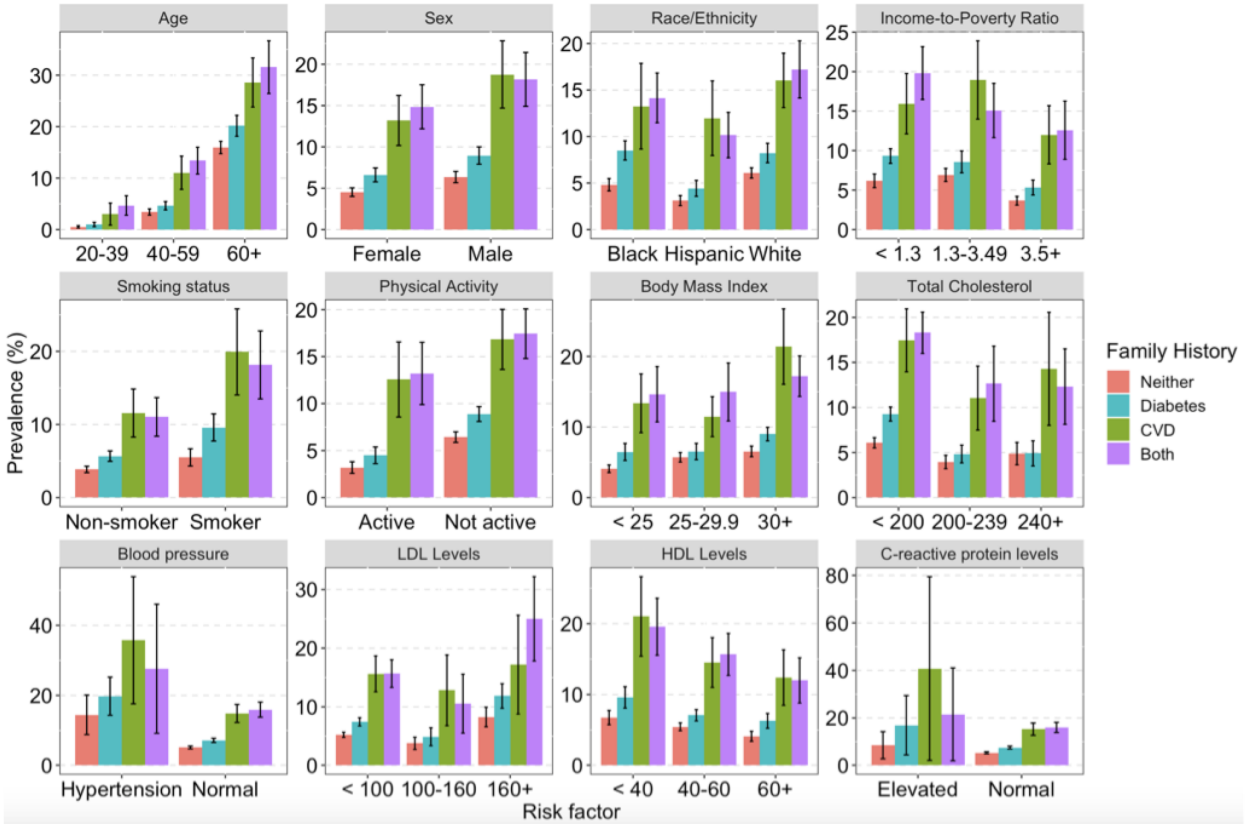
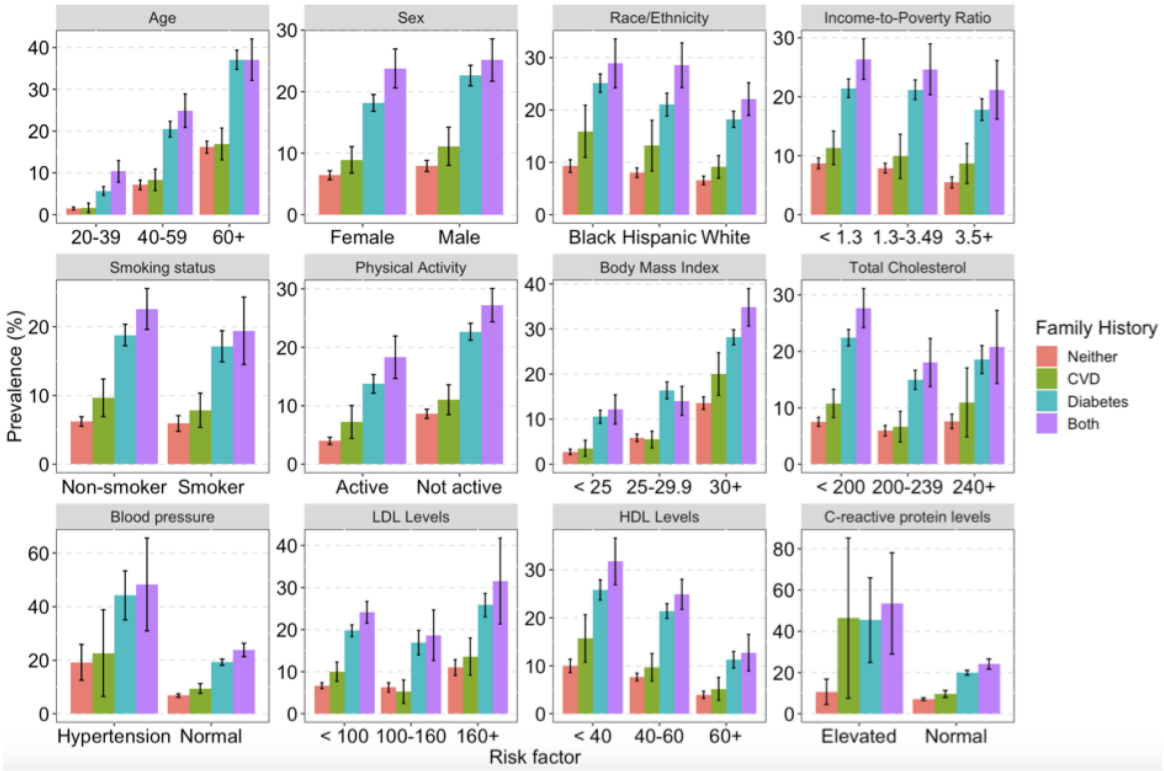
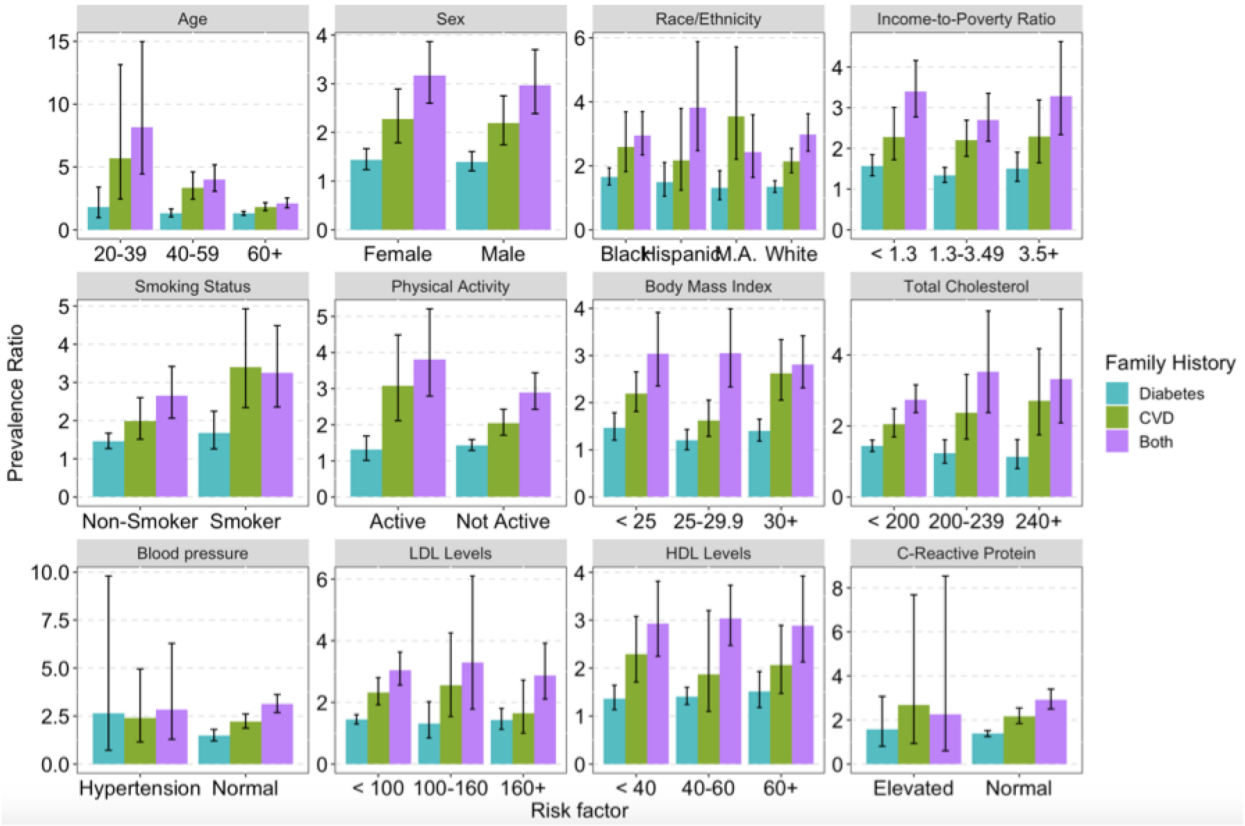


Figure 2.3a. Survey-adjusted prevalence of CVD within cohorts of individuals by selected risk factors and family history of neither CVD nor diabetes, diabetes only, CVD only, and both CVD and diabetes, NHANES 2007-2018. Note: y-axes for multiple panels have different ranges.

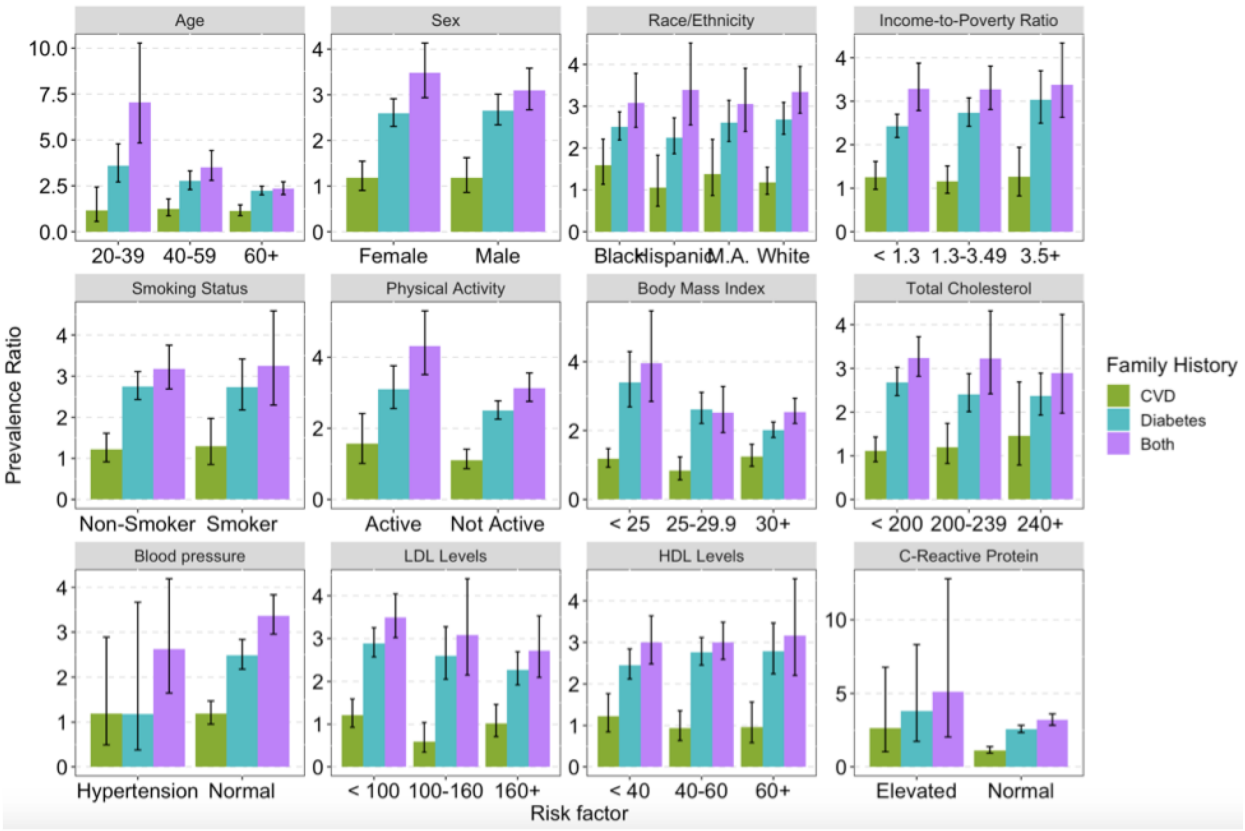


[1] Race/ethnicity groups labeled as "black" and "white" are non-Hispanic

Figure 2.3b. Survey-adjusted prevalence of diabetes within cohorts of individuals by selected risk factors and family history of neither CVD nor diabetes, diabetes only, CVD only, and both CVD and diabetes, NHANES 2007-2018. Note: y-axes for multiple panels have different ranges.



Supplementary Figure 2.4a. Prevalence ratio for CVD according to family history of diabetes, CVD, and both, adjusted by age, sex, and race/ethnicity, in sub-cohorts of individuals with selected demographic characteristics and CVD/diabetes risk factors. Reference group is cohort of individuals with no family history of CVD and diabetes.



[1]Covariates in multivariable model include age, sex, and race/ethnicity

Supplementary Figure 2.4b. Prevalence ratio for CVD according to family history of diabetes, CVD, and both, adjusted by age, sex, and race/ethnicity, in sub-cohorts of individuals with selected demographic characteristics and CVD/diabetes risk factors. Reference group is cohort of individuals with no family history of CVD and diabetes

Family history associated with increased prevalence of CVD/diabetes risk factors in participants not reporting diabetes or heart disease

Among participants without prevalent CVD or diabetes and who did not have abnormal glucose or hemoglobin A1C levels, a positive family history of CVD and/or diabetes was significantly associated with CVD and diabetes risk factors. A positive family history of both CVD and diabetes was significantly associated with a BMI greater than 30 (aOR 1.70; 95% CI, 1.48-1.94), a poverty-income-ratio less than 1.3 (aOR of 1.51; 95% CI,

1.31-1.73), smoking (aOR 2.15; 95% CI, 1.81-1.56), and low HDL levels (aOR 1.36; 95% CI, 1.20-1.54) (**Table 2.3**).

Table 2.3. Adjusted odds ratios (95% CI) for CVD/diabetes risk factors according to family history, adjusted by age, sex, and race/ethnicity, in individuals without diabetes and CVD, NHANES 2007-2018.

Reported family history of:	CVD only	Diabetes only	CVD and Diabetes
BMI kg/m2 (<25†)			
25-29.9	1.10 (0.95,1.26)	0.95 (0.89,1.03)	0.90 (0.79,1.03)
30+	0.92 (0.79,1.07)	1.47 (1.36,1.58)*	1.70 (1.48,1.94)*
Poverty-Income-ratio (≥3.5†)			
<1.3	1.12 (0.96,1.30)	1.02 (0.95,1.09)	1.51 (1.31,1.73)*
1.3-3.49	1.12 (0.97,1.29)	0.98 (0.91,1.06)	1.15 (0.98,1.34)
Education (≥high school†)			
< high school	1.41 (1.19,1.68)*	0.89 (0.82,0.97)*	1.55 (1.32,1.83)*
Current smoker (No†)			
Yes	1.51 (1.28,1.80)*	1.15 (1.04,1.26)*	2.15 (1.81,2.56)*
Physically active (Yes†)			
No	1.17 (1.00,1.36)	0.96 (0.88,1.05)	1.25 (1.07,1.44)*
Total cholesterol (Not high†)			
200-239	0.81 (0.67,0.97)	1.06 (0.98,1.14)	1.06 (0.92,1.23)
240+	1.00 (0.80-1.25)	1.08 (0.86,1.35)	1.08 (0.86-1.35)
Blood pressure (Normal†)			
Hypertension	0.81 (0.54,1.24)	0.81 (0.62,1.04)	0.97 (0.59,1.61)
CRP levels (Not high†)			
High	0.85 (0.58,1.25)	1.25 (1.03,1.52)	1.45 (1.09,1.94)*
LDL levels (Not high†)			
Borderline high	1.13 (0.94,1.35)	0.89 (0.81,0.99)	0.91 (0.74,1.1)
High	1.24 (1.02,1.51)	0.86 (0.77,0.96)*	0.90 (0.70,1.15)
HDL levels (High†)			
Low	0.94 (0.79,1.12)	1.14 (1.06,1.23)*	1.36 (1.20,1.54)*

*Values marked with an asterisk meet an FDR significance threshold of 5%.

Earlier age of diabetes diagnosis for participants with family histories of both CVD and diabetes

For participants diagnosed with diabetes, the weighted mean age of diabetes diagnosis for participants with family histories of both CVD and diabetes was 44.07 years, or on average, 6.62 years earlier ($P < 0.0001$) than participants with no family history of CVD and diabetes. Furthermore, for participants diagnosed with diabetes, participants with a family history of diabetes and no family history of CVD were diagnosed with diabetes, on average, at 48.16 years, or 2.53 years earlier ($P < 0.0001$) than participants with no family history of CVD or diabetes and participants with a family history of CVD and no family history of diabetes were diagnosed with diabetes, on average, at 51.31 years, or 0.62 years later ($P < 0.0001$) than participants with no family history of CVD or diabetes (data not shown).

DISCUSSION

Using a large population-based survey, we document the combined impact of family history of CVD and diabetes on the population prevalence of CVD, diabetes, and CVD and diabetes risk factors. First, we show that over the period 2007-2018, an average 44.4% of the US adult population has a family history of CVD or diabetes and that having one is associated with the other. Specifically, 5.5%, 31.6%, and 7.3% of the US adult population aged 20 years and older has a family history of CVD only, diabetes only, or family histories of both, respectively. Second, our findings suggest that prevalence of diagnosed CVD and/or diabetes among a population with a positive family history is greater than that of a population with no family history, independent of established risk factors, such as BMI, age, and sex. Specifically, for participants with no family history,

the prevalence of CVD, diabetes, and both CVD and diabetes is 5.4%, 7.1%, and 1.3%, while for participants with both family histories of CVD and diabetes, the prevalence increases to 16.0%, 24.3%, and 7.4%, respectively. Third, we show how family history associated prevalence and PRs for CVD and diabetes remain similar across low/desirable and high levels of certain CVD/diabetes risk factors, such as LDL cholesterol, total cholesterol, and blood pressure as shown in Figure 2 and Supplementary Figure 3. Further, we find that participants with family histories of both CVD and diabetes are diagnosed with diabetes, on average, 6.6 years earlier than participants without either family history, indicating the potential utility of collecting family history information for early detection[44,45].

Epidemiological studies have attempted to disentangle the complex relationship between CVD and diabetes, in search of common pathophysiological pathways and potential mechanistic roles linking the two diseases[46]. Many factors are shared between both diseases, including adiposity and dyslipidemia, cardiometabolic markers and inflammatory profiles[47,48]. While epidemiological studies have established a higher cardiovascular and atherosclerotic burden in participants with diabetes, there is also evidence for vascular abnormalities being present prior to diabetes onset, leading to the "common soil" hypothesis that both arise from a shared antecedent[25,49–51]. Studies examining the shared genetic architecture of CVD and diabetes have uncovered overlapping genetic loci from large genome-wide association studies (GWAS)[52,53]. Our study builds on these findings by leveraging family history information, which is reflective of both total shared genetic variations (e.g. GWAS single nucleotide

polymorphisms (SNPs) and other sources of genetic variation) at multiple loci as well as shared environmental, behavioral and lifestyle factors. By demonstrating the high prevalence of CVD and diabetes associated with family history, we highlight the importance of collecting family history information for early detection and prevention of these two chronic conditions.

This study has several limitations. First, NHANES is a cross-sectional survey and therefore cannot be used to establish causal relationships between family history, risk factors, and diagnosis of CVD and diabetes. Second, we used self-reported family history and disease diagnosis information, which is prone to recall biases that may contribute to measurement error. Although the NHANES family history questionnaire specifically mentioned parents and siblings as first-degree relatives, we cannot rule out the possibility that participants included second-degree and surrogate relatives in their reporting of family history. However, previous studies examining the accuracy of reported family history to that of physician-assessed status of close relatives or status reported by parents have found a sensitivity and specificity of 78.5% and 94.9% for family history of diabetes and 85% and 93% for family history of coronary heart disease, which may attenuate the magnitude of our associations[18,19]. Furthermore, studies have found the sensitivity of self-reported diabetes and heart disease were 96% and 78%, respectively, and specificity ranged from 95% to 99%[16,17]. Although we were not comprehensive in our definition of CVD due to NHANES self-report instrument limitations, we included participants with coronary heart disease, angina, heart attack, and stroke to ensure we identified participants with symptomatic indicators of cardiovascular disease. A third limitation is our inability to discriminate between type 1 and type 2 diabetes because the NHANES

questionnaire does not ask participants to specify. However, approximately 95% of diagnosed diabetes cases in adults are type 2 diabetes[54].

Public health efforts in the prevention of common chronic diseases tend to be disease-oriented. For example, the identification of individuals at high risk of diabetes typically starts with asking about family history of diabetes[55]. On the other hand, efforts to reduce the burden of heart disease have used family history of heart disease[30]. Clinical guidelines for standards of medical care published by the American Diabetes Association screen for a family history of type 2 diabetes in first and second-degree relatives; however, given our findings on the increased risk individuals with both family histories of CVD and diabetes have, we suggest that information on both family histories be used as a screening tool for identifying high risk individuals.

Family health history as a risk factor cuts across multiple chronic diseases and is important for evaluating health risks of many diseases. Millions of people in the United States have a family history of CVD and/or diabetes, putting them at increased risk for one or both of these conditions. The strong association between family history and prevalence of CVD and/or diabetes revealed by our study warrants further investigation into the genetic and environmental factors that compose family history, beyond the common risk factors considered here. Through public educational offerings, the CDC hosts the Surgeon General's My Family Health Portrait, an online tool for collecting family health history for multiple disease conditions, with a focus on heart disease, diabetes, and cancer[56]. Our findings underscore the public health impact and utility of

family history as a tool for screening and identifying high-risk populations across chronic diseases.

We extend this work in *Chapter 3*, where we scale up this technique to examine all pairwise family history of disease in multiple clinical outcomes, revealing undiscovered shared genetic architecture and environmental influences of seemingly unrelated diseases.

3

Genetic and Environmental Components of Cross-Disease Familial Risk (XY-FamWAS)

Estimates of cross-disease genetic and environmental components of familial risk in *UK Biobank* ("XY-FamWAS"): Is disease X associated with a family history of a different disease Y?

Collaborators: Yixuan He, Chirag Lakhani, Arjun Manrai, Chirag J Patel

ABSTRACT

Background. Most common chronic diseases, such as type 2 diabetes and cancer, result from the complex interplay between genetic factors and shared environmental exposures, reflected in family history of disease. While the increased odds for a disease due to a family history has been documented for many chronic conditions, the increased odds of a disease X associated with a family history of a different disease Y has not been

established, and could implicate shared genetic and environmental etiologies between diseases.

Methods. We investigated 132 cross disease-family history associations for 12 complex human diseases in up to 500,000 participants from the UK Biobank. Next, we performed a range of analyses to assess the components of familial risk attributed to genetic or environmental factors. First, we identify shared genetic architecture of cross disease-family history by estimating genetic correlation using linkage disequilibrium score regression. Second, we decompose the shared environment by comparing cross disease-family history associations in a cohort of 6,347 adopted individuals (that share zero genetics) to those of non-adopted individuals. Third, we dissect the influence of genotypic factors through a genetic risk score incorporated for disease Y in investigating the association between disease X associated with a family history of a different disease Y.

Results. In our main findings, we recapitulate the observational associations between the same family history and disease pairs (e.g., family history of diabetes and diabetes: OR=3.49). We broadened our search to 132 non-same pairs of family history and disease, scaling up the search to 12 possible diseases. In this search, we found unexpected associations, including between family history of emphysema/chronic bronchitis and depression (OR 1.30; 95%CI 1.23-1.37, FDR < 0.05), which was driven more by maternal than paternal history (OR 1.36 vs OR 1.20 versus no family history). The association exhibited a weak genetic correlation (rg = 0.15 SE = 0.24) and a stronger

magnitude of association in an adopted cohort (OR 1.4; 95%CI 1.11-1.78), reflecting the contribution of shared environmental factors for depression in families with emphysema/chronic bronchitis.

Conclusion. Our atlas of disease and family history associations demonstrate the shared genetic architecture and environmental factors underlying many seemingly dissimilar complex diseases.

INTRODUCTION

Many common complex conditions such as coronary heart disease, diabetes, and cancer, share genetic risk variants and environmental exposure risk factors. As we have shown in the previous chapters, family history of disease captures genetic susceptibility as well as the environmental, lifestyle, and behavioral habits shared within families whereby an individual may be predisposed to be at higher risk for disease. Leveraging data on phenotyped relatives with missing genotypes can reveal undiscovered genetic factors or environmental influences associated with disease [57]. Discovering shared risk factors is of clinical importance because it can provide insight into potential shared genetic architecture among diseases, highlight shared disease mechanisms, and pinpoint potential areas for prevention or therapeutic intervention. In this study, we leverage family history to discover potential shared familial influences between 12 complex human diseases.

Further, knowledge of whether family history associated disease risk is driven by shared genetic variation or environmental and behavioral factors can improve early detection

and prevention strategies. A classical genetics tool for disentangling genetic influences from environmental factors is an adoption study, where the outcome of an adopted individual (who shares zero genetics with relatives) is compared to that of a non-adopted individual. Recent genetic methods such as estimation of genetic correlation via linkage disequilibrium (LD) score regression or polygenic risk score (PRS) analysis take into account the associations between multiple SNPs within the genome and can provide a more comprehensive analysis of shared genetic architecture of complex diseases and genetically characterized clinical risk. For example, while observational studies have shown an association between Alzheimer's Disease (AD) and an increased risk in ischaemic stroke, genome-wide association studies have been unsuccessful in identifying genome-wide significant SNPs shared among both diseases, indicative of a shared non-genetic etiology or study bias. Complex diseases such as AD and stroke, can, however, be driven by joint interactions of many associated genes that fall below genome-wide significance threshold or share non-genetic, but familially transmitted factors [58]. Applying these methods to examine "cross" disease-family history associations can illuminate shared genetic and non-genetic (environmental) familial effects across seemingly different diseases and conditions. Dissecting the genetic and shared environmental components of disease can group diseases based on genetic-only and environmental-only classifications and reconstruct complete phenotypic classifications [59].

In the first part of our study, we perform 132 cross disease-family history associations for every disease X and family history of disease Y('XY-FamWAS') in up to 500,000

participants of the UK Biobank, a large population-based cohort in the United Kingdom. In the second part of our study, we attempt to deconvolve the genetic and environmental components of cross disease-family history associations using three different methods. First, we examine pairwise disease-family history associations in an adopted cohort to isolate non-genetic phenomena. Second, we perform genetic correlation analyses between disease and family history to uncover their shared genetic architecture, if at all. Third, in order to examine how much of the association can be explained by PRS-defined genetics, we incorporate a PRS for disease Y in investigating the association between disease X and family history of a different disease Y. Our findings discover novel disease-family history associations, thereby uncovering putative overlap of shared genetic architecture and environmental influences by which seemingly different diseases are related.

METHODS

Study Population

The UK Biobank (UKB) is a national health resource consisting of 502,628 participants (aged 40-69 years) recruited between 2006 and 2010 from the general population of the United Kingdom, spanning England, Wales, and Scotland [60]. The resource has collected detailed data on indicators of the health status of its participants through extensive questionnaire assessments including demographic and medical information, physical measurements, genome-wide genotyping, and quantification of environmental exposures through blood and urine samples [61].

Family history assessment

Participants were asked on a touchscreen questionnaire whether any first degree biological and adopted relatives (father, mother, and siblings) had any common serious illnesses. Specifically, biological family history was ascertained with the following questions: "Has/did your father ever suffer from?", "Has/did your mother ever suffer from?", and "Have any of your brothers or sisters suffered from any of the following diseases?". Adopted family history was ascertained with the following questions: "Has/did your adopted father ever suffer from?" and "Has/did your adopted mother ever suffer from?". The sets of illnesses that participants were asked to answer included: heart disease, stroke, high blood pressure, chronic bronchitis/emphysema, Alzheimer's disease/dementia, diabetes, Parkinson's disease, severe depression, lung cancer, bowel cancer, prostate cancer, and breast cancer.

Disease Phenotypes

We ascertained participant current disease status for every family history phenotype. The data-fields used to ascertain current disease status are presented in **Supplementary Table 3.1**. Specifically, we ascertained self-reported vascular and heart problems by the following self-reported question: "Has a doctor ever told you have any of the following conditions?". The possible touchscreen responses were: heart attack, angina, stroke, high blood pressure, none of the above, and prefer not to answer, where participants were asked in an initial assessment visit (2006-2010), a repeat visit (2012-2013), and an imaging visit (2014+). We ascertained participants with diabetes by an affirmative response to the question, "Has a doctor ever told you that you have diabetes?". We

ascertained participants with a non-cancer illness, including depression, Parkinson's disease, Alzheimer's disease, emphysema/chronic bronchitis, and stroke, and a self-reported cancer diagnosis, including prostate cancer, breast cancer, bowel cancer, and lung cancer, by an affirmative response to a diagnosis. To maintain accuracy of data collection, all affirmative medical conditions questionnaire responses on the UK Biobank touchscreen were then asked directly to the participant by a study nurse.

Statistical analyses for observational associations

We used logistic regression to test the association between every pair of family history of disease and prevalent disease, adjusting by age, sex, and 15 genetic principal components (PCs) derived from genotype data. We defined family history as a binary variable (coded as "0" or "1") for any first-degree relative (father, mother, sibling) with disease. We conducted separate cross disease-family history analyses in adopted (n = 6347) versus non-adopted adults to disentangle genetic effects from environmental influences. We additionally conducted separate analyses to test the association between paternal, maternal, and sibling history with prevalent disease in order to detect differences in association for different types of first-degree relatives (mother, father, sibling). We corrected for multiple testing across all pairs of tests using the false discovery rate (FDR) method [10]. We report findings that are deemed to be significant at a P-value less than 0.05 and at a more stringent threshold of FDR of 5%.

To determine the predictive capability of family histories of different diseases Y associated with a disease X, we constructed logistic regression models by regressing each

predictor set on current disease status. Predictor sets consisted of combinations of: age and sex (covariates), family history of disease X, and family histories of diseases Y identified to be significantly associated at an FDR threshold of 5% with disease X and exclusive of family history of disease X itself. We calculated the area under the receiver operating characteristic curve (AUROC, a common measure of predictive power, see [62]) with the pROC package in R, with AUROC 95% confidence intervals computed across all bootstraps [63]. Individuals with missing values were removed in order to perform predictions from a fitted generalized linear model (GLM).

Genome-Wide Association Study and Linkage Disequilibrium Score regression

Next, we sought to describe the shared genetic architecture between complex disease and family history. We conducted GWAS analyses on family history phenotypes using PLINK, where the regression models for family history were adjusted for age, sex, and 15 genetic principal components. We conducted quality control steps, including removing individuals of non-British ancestry. Due to limited sample sizes, we were only able to perform a GWAS on 7 of the 12 family history phenotypes, including: breast cancer, bowel cancer, diabetes, emphysema/chronic bronchitis, Parkinson's Disease, prostate cancer, and stroke; and 9 of the 12 disease phenotypes, including: Alzheimer's Disease, breast cancer, bowel cancer, depression, diabetes, emphysema/chronic bronchitis, heart disease, prostate cancer, and stroke. We then applied Linkage Disequilibrium Score (LDSC) regression to estimate the genetic correlation between family history and disease by regressing association test statistics for SNPs on their LD scores. We used the LDSC tool by Bulik-Sullivan et al. for performing these analyses [15,64].

Polygenic Risk Score for breast cancer, type 2 diabetes, and coronary artery disease

Polygenic risk scores (PRS) provide a quantitative measure summarizing the cumulative effects of a number of risk alleles based on an individual's genotype. We leveraged PRS in order to identify how much of the association between disease X and family history Y can be explained by a genetic risk profile for disease Y. We performed PRS analyses using scores derived by Khera et al. [65] for breast cancer, type 2 diabetes, and coronary artery disease. To compute each individual's PRS score, we computed the sum of the product of the number of risk alleles each individual carries by the weight assigned to the genetic variant representing the strength of the association of the variant with disease risk (i.e., log of the odds ratio of the allele), across all identified SNPs. For each of these three disease phenotypes, we used logistic regression analyses to measure the association of disease X (coded as binary '0' or '1') with family history of disease Y, adjusting for the PRS for disease Y, age, sex, and 15 principal components. We report findings that are deemed to be significant at a P-value less than 0.05.

RESULTS

Participant characteristics

Demographic information for the study cohorts are provided (**Supplementary Table 3.2**). The average age of participants was 56.5 and 56.2 for the non-adopted and adopted cohorts, respectively. The breakdown and prevalence of self-reported disease and family history of disease for each cohort is provided in **Supplementary Table 3.3**. High blood

pressure (27.58% of participants), depression (5.96%), and diabetes (5.39%) were the most prevalent diseases among the non-adopted cohort, as well as in the adopted cohort (29.52%, 7.56%, 7.96%, respectively). Family history of cardiometabolic traits were the most prevalent of all family histories in the non-adopted cohort, including family history of high blood pressure (48.27%), heart disease (43.45%), stroke (26.64%), and diabetes (21.90%), as well as in the adopted cohort (31.72%, 34.09%, 20.39%, 13.80%). The cohort breakdown and prevalence of maternal, paternal, and sibling family history is shown in **Supplementary Table 3.4**.

Family history *Y* as a risk factor for another disease *X*

Figure 3.1 presents all 144 pairwise associations between 12 complex human diseases and their family histories. For each association, individuals with a positive family history and who had the disease of the family history were removed from analyses. The estimates of family history associations with disease for same disease pairs are consistent [2], including diabetes (OR 3.49, 95%CI 3.40-3.58; FDR-adjusted p-value [FDR] < 0.001), prostate cancer (OR 3.03, 95%CI 2.79-3.29; FDR < 0.001), breast cancer (OR 1.95, 95%CI 1.86-2.05; FDR < 0.001), stroke (OR 1.44, 95%CI 1.37-1.51; FDR < 0.001), and heart disease (OR 2.55, 95%CI 2.47-2.64; FDR < 0.001) (**Supplementary Figure 3.1a**). In other words, for example, individuals with a family history of diabetes had a 3.5 fold increased odds for having prevalent diabetes (Supplementary Figure 1a), of which less than 2% of the association could be attributed to PCs (**Supplementary Figure 3.1b, 2, 3**). Overall, we identified a higher magnitude of association for same disease-family

history pairs ($OR_{AVG} = 2.45$) versus different disease-family history pairs ($OR_{AVG} = 1.11$), for findings significant at a P-value < 0.05.

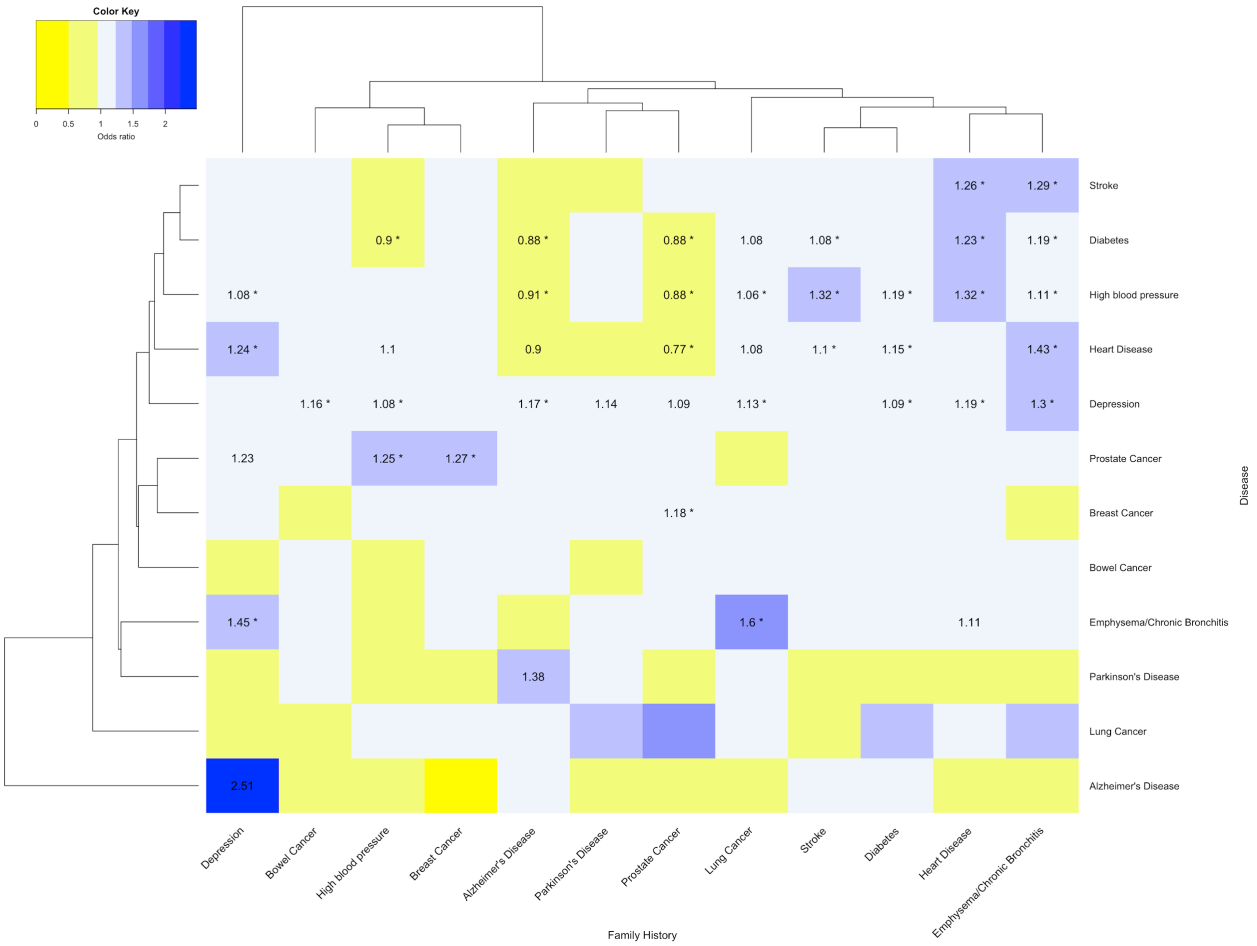


Figure 3.1. All 144 pairwise associations between 12 complex human diseases and their family histories. For each association, individuals with a positive family history and who had the disease of the family history were removed from analyses. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P* value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age, sex, and 15 principal components.

A family history of the cardiometabolic diseases (stroke, diabetes, heart disease) was positively associated with other cardiometabolic conditions (OR 1.09-1.32; FDR < 0.05) (stroke, diabetes, high blood pressure, heart disease). For example, we identified a

positive association between prostate cancer and family history of breast cancer (OR 1.27, 95%CI 1.06-1.51; FDR < 0.05) as well as breast cancer and family history of prostate cancer (OR 1.18, 95%CI 1.04-1.33; FDR < 0.05). We also identified a strong association between emphysema/chronic bronchitis and a family history of lung cancer (OR 1.60; 95%CI 1.46-1.76; FDR < 0.001), and with a family history of heart disease (OR 1.11, 95%CI 1.02-1.21; FDR < 0.05) (**Figure 3.1**).

We also found some associations that are, to the best of our knowledge, not previously reported in epidemiological or genetic studies, or remain controversial. First, we found a protective effect (negative association) of family history of prostate cancer on heart disease (OR 0.77, 95%CI 0.69-0.87; FDR < 0.001), high blood pressure (OR 0.88, 95%CI 0.84-0.92; FDR < 0.001), and diabetes (OR 0.88, 95%CI 0.80-0.96; FDR < 0.05) (**Figure 3.1**). Second, we identified positive associations with the condition of depression, including family history of bowel cancer, high blood pressure, Alzheimer's disease, Parkinson's disease, prostate cancer, lung cancer, diabetes, heart disease, and emphysema/chronic bronchitis (**Figure 3.1**). Third, we identified an association between the prostate cancer condition and a family history of depression (OR 1.23, 95%CI 1.01-1.49; p < 0.05). We identified minimal differences in comparing disease-family history associations for male versus female participants (**Supplementary Figure 3.4**).

Further, a prediction model for emphysema/chronic bronchitis that included family histories of diabetes, depression, emphysema, heart disease, lung cancer, and covariates, and not family history of emphysema/chronic bronchitis itself, achieved the same area

under the receiver operating characteristic (ROC) curve (AUC) than that achieved by a family history of emphysema/chronic bronchitis (AUC 0.70 vs 0.69) (**Supplementary Figure 3.5**).

Maternal history more strongly associated with disease outcome than a paternal history, and sibling history is more strongly associated than maternal history

We documented the maternal, paternal, and sibling components of family history (Supplementary Figure 6). For clarity of presentation, and in order to compare maternal to paternal history contribution to family history associated risk, we plotted only associations that we had evidence were non-zero (before multiple comparisons) by choosing associations that were significant at a p-value less than 0.05 in both parental associations (**Figure 3.2, Continued**). Overall, we identified a higher magnitude of association for maternal ($OR_{AVG} = 1.27$) versus paternal history ($OR_{AVG} = 1.19$), for findings significant at a P-value < 0.05(**Figure 3.2a, Continued**). Compared to an overall magnitude of association of 1.4 for any (maternal, paternal, or sibling) family history, we identified an average magnitude of association of 1.29 and 1.55 for maternal-only and sibling-only history, respectively (**Figure 3.2b and c, Continued**).

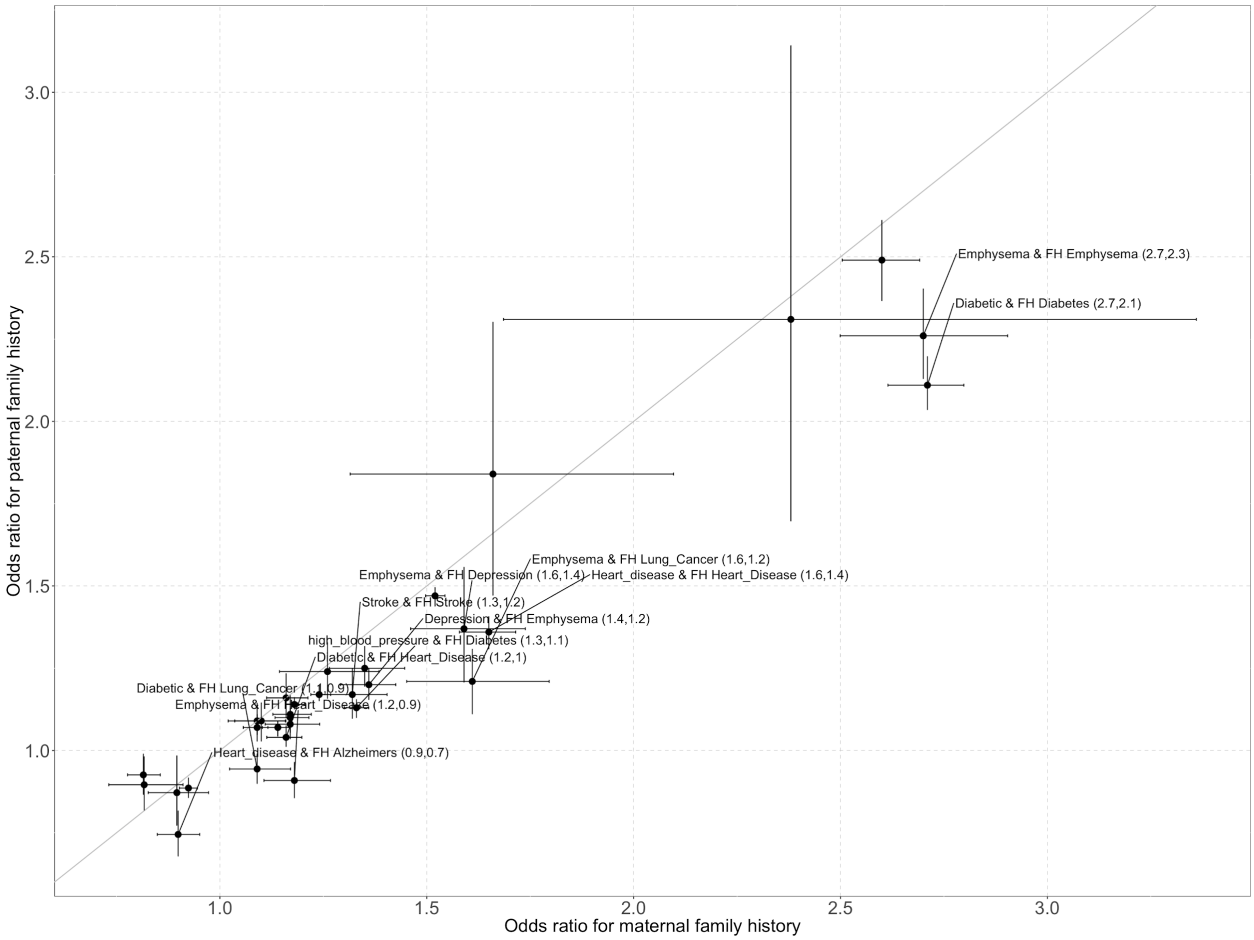
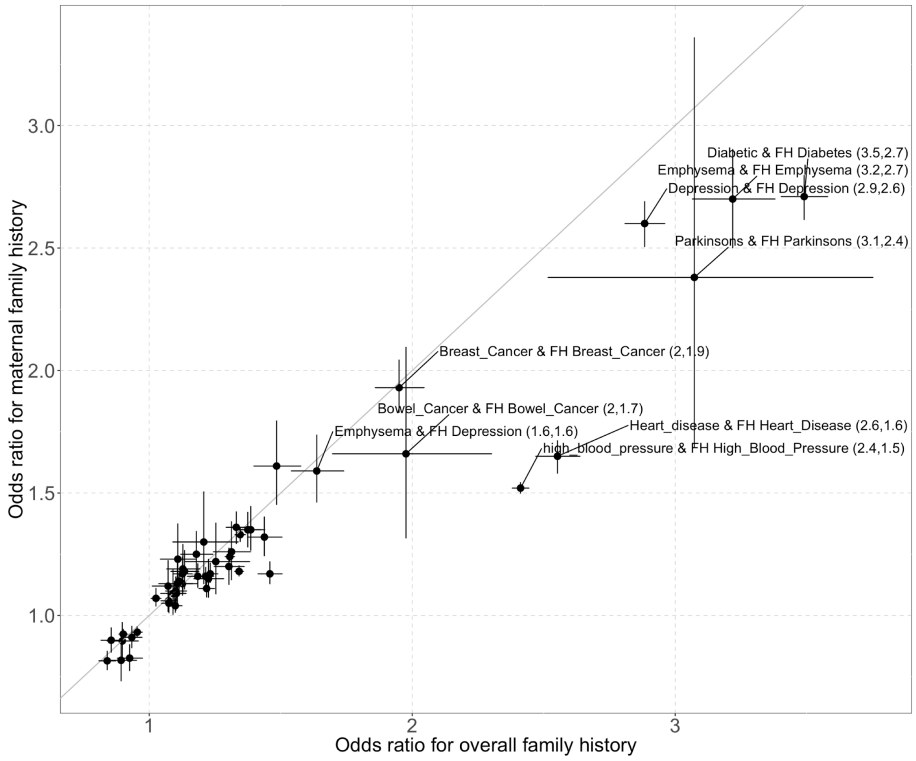


Figure 3.2. (A) Maternal versus paternal family history. Disease-family history associations for maternal history (x-axis) are presented against associations for paternal history (y-axis). Horizontal and vertical error bars represent 95% confidence intervals. All points represented are significant at P-value < 0.05 in both maternal and paternal analyses. Each association is annotated with a label listing the disease followed by family history (FH) and a numeric label in parentheses listing the odds ratios for maternal and paternal history.



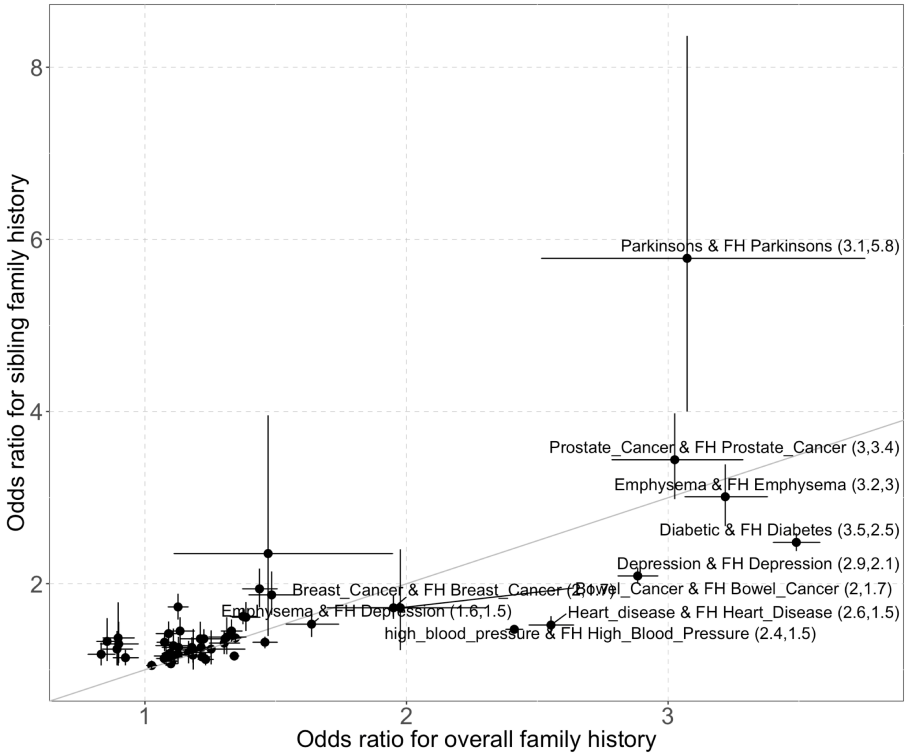


Figure 3.2. (B) Overall family history versus maternal-only family history. **(C)** Overall family history versus sibling-only family history.

The largest and smallest magnitude of maternal association was diabetes and a family history of diabetes (OR = 2.71), and stroke and a family history of Parkinson's Disease (OR = 0.796) and the largest and smallest magnitude of paternal association was heart disease and family history of Alzheimer's Disease (OR = 0.745) and prostate cancer and family history of prostate cancer (OR = 2.65) (**Supplementary Figure 3.6a**).

The paternal magnitude of association for diabetes and a family history of diabetes was less than the maternal (OR = 2.11), and for emphysema/chronic bronchitis and a family history of lung cancer (OR = 1.21) (**Supplementary Figure 3.6b**). Further, we identified a positive maternal association and a negative paternal association with diabetes and a family history of lung cancer (OR 1.09 vs. 0.94) and emphysema/chronic bronchitis and family history of heart disease (OR 1.18 vs. 0.91).

Sibling history, on the other hand, achieved a higher magnitude of association than both maternal and paternal history ($OR_{AVG} = 1.58$), where we identified the largest and smallest magnitude of association for Parkinson's Disease and a sibling history of Parkinson's Disease (OR $= 5.78$), and high blood pressure associated with lung cancer (OR $= 1.05$) (**Supplementary Figure 3.6c**).

Family history of adopted participants highlights the role of shared environment for high blood pressure, emphysema/chronic bronchitis, and depression

To disentangle potential environmental contributions from genetic influences, we compared disease-family history associations in the adopted cohort to those found in the

non-adopted cohort (**Figure 3.3 [continued], Supplementary Figure 3.7**). The overall magnitude of association in the adopted cohort was 1.63, compared to OR 1.92 in the non-adopted cohort, for findings significant at a P-value < 0.05 (**Figure 3.3, continued**). We identified a greater magnitude of association in the adopted versus non-adopted cohort for depression and family history of stroke (OR 1.4 vs 1.1) and emphysema and family history of lung cancer (OR 1.9 vs. 1.5). We noticed that for same disease-family history pairs, the effects were consistently larger in the non-adopted cohort for depression, heart disease, emphysema, diabetes, and high blood pressure, informing on the genetic influence of familial risk for these cardiometabolic traits. Further, in order to determine how much of the disease-family history association can be explained by shared non-genetic or environmental factors, we additionally adjusted our models by current smoking status and income. Family history remained a significant independent risk factor. We found that all (n=12) but 1 association (emphysema and family history of lung cancer, p-value = 0.09, B1 = 0.62, B1adj = 0.45) remained significant after adjusting for these shared non-genetic or environmental factors. We found that, on average, smoking and income represented 0.07% of the disease-family history associations, where the highest percent difference was for the association between heart disease and adopted family history of heart disease (16%).

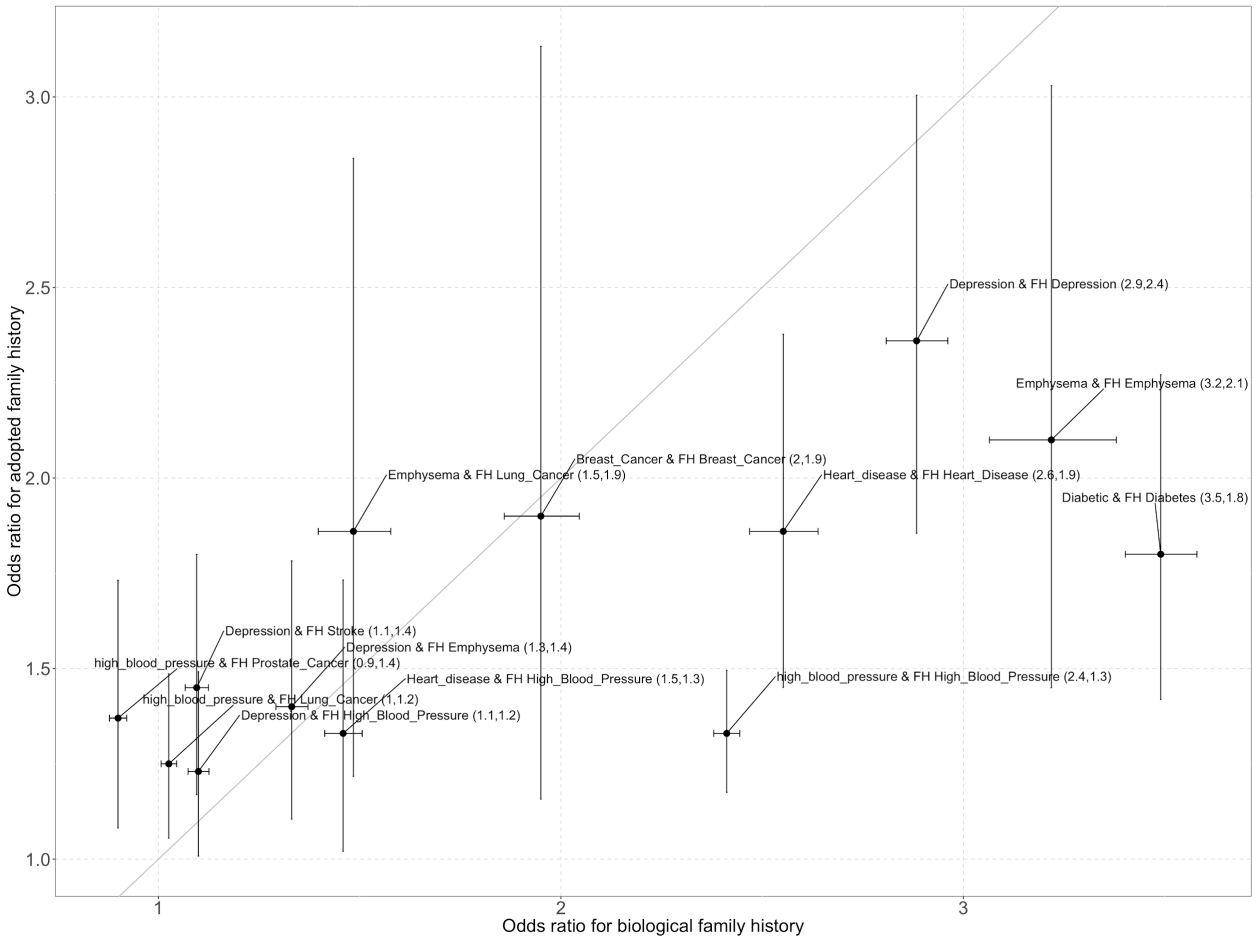


Figure 3.3. Non-adopted compared to adopted family history. Disease-family history associations for non-adopted (biological) family history (x-axis) are presented against associations for adopted family history (y-axis). Horizontal and vertical error bars represent 95% confidence intervals. All points represented are significant at p-value < 0.05 in both non-adopted and adopted analyses. Each association is annotated with a label listing the disease followed by family history (FH) and a numeric label in parentheses listing the odds ratios for non-adopted (biological) and adopted history.

Next, we attempted to separate maternal and paternal adopted family history influences (**Supplementary Figure 3.8, 3.9**). We identified, for example, greater magnitudes of association for depression given adopted maternal history of high blood pressure, stroke, and heart disease compared to non-adopted maternal history (**Supplementary Figure 3.8, 3.9**).

Further, we noticed the clustering of disease and family history of disease phenotypes based on magnitudes of association were different in the adopted analysis (**Supplementary Figure 3.7**) to that of the non-adopted (**Figure 3.1**), including the clustering of family history of emphysema/chronic bronchitis, breast cancer, and prostate cancer in the adopted cohort. We identified emphysema/chronic bronchitis clustered among the cardiometabolic traits in the adopted cohort, as well as a family history of Alzheimer's disease and diabetes and a family history of lung cancer and breast cancer, which was not the case for the non-adopted cohort (**Supplementary Figure 3.7**).

Common variants and genetic predisposition for BC, CAD, and T2D

To further tease apart the contribution of additive and common genetic variants in disease, we estimated the associations between family history of disease and disease adjusted by a polygenic risk score. The association of family history with disease, adjusting for polygenic risk score, as well as age, sex, and principal components (covariates) for breast cancer, diabetes, and heart disease are presented in **Table 3.1**. First, among the same disease-family history pairs, we found the disease associated with a family history of the disease, even after adjusted for PRS (P_{FH} < 2e-16 and P_{GRS} < 2e-16 for BC, CAD, and T2D). We found PRS represented 7.02%, 7.02%, and 12.78% of the same pair disease-family history associations for diabetes, heart disease, and breast cancer, respectively (**Table 3.1, continued**).

Table 3.1. Common variants and genetic predisposition for breast cancer, heart disease, and type 2 diabetes explain a small proportion of disease-family history associations.

Family History	Self-reported disease	P-value for family history [Model 2]	P-value for PRS [Model 2]	B for PRS [Model 2]	B1adj [Model 2]	B1 [Model 1]	% change (B1adj-B1)/B1
Diabetes	Diabetes	< 2e-16	< 2e-16	3.812	1.126	1.2118778	-7.02%
Diabetes	Heart Disease	5.13e-14	< 2e-16	1.202	0.2517	0.2845	-11.53%
Diabetes	Breast Cancer	0.6274	0.8715	0.0260486	-0.0214109	-0.0207150	-3.36%
Heart Disease	Diabetes	< 2e-16	< 2e-16	1.167	0.2454	0.2640293	-7.06%
Heart Disease	Heart Disease	< 2e-16	< 2e-16	4.599	0.8891	0.9562	-7.02%
Heart Disease	Breast Cancer	0.8654	0.9228	-0.0205779	0.0061706	0.0058317	5.81%
Breast Cancer	Diabetes	0.56172	0.50796	0.0030305	-0.0216080	-0.0201969	-6.99%
Breast Cancer	Heart Disease	0.644137	0.259498	-0.0064442	0.0211303	0.01813	16.55%
Breast Cancer	Breast Cancer	< 2e-16	< 2e-16	0.181	0.5662	0.6492	-12.78%

* For every disease X and family history of a different disease Y represented as $FH(Y)$, $B1$ is derived by the logistic regression model $X \sim B_1 FH(Y) + age + sex + PCs$ [Model 1].

** $B1adj$ is derived by the logistic regression model $X \sim B_{1adj} FH(Y) + PRS(Y) + age + sex + PCs$ [Model 2].

Second, among non-same disease-family history pairs, for diabetes and heart disease, we found heart disease associated with a family history of T2D, even after adjusting for T2D PRS (P_{FH} < 2e-16 and P_{GRS} < 2e-16), and the same held for diabetes and family history of heart disease (P_{FH} < 2e-16 and P_{GRS} < 2e-16). We found family history of T2D was strongly associated with prevalent CAD (OR 1.29), adjusting for T2D PRS and covariates, where 11.53% of the association was explained by a T2D PRS, indicating that the remaining 88.5% is due to genetics not captured by a polygenic risk score and/or environmental factors. Further, we found diabetes associated with a family history of CAD (OR 1.28), even after adjusting for CAD PRS (P_{FH} < 2e-16 and P_{GRS} < 2e-16), which explained 7.06% of the association. The association of diabetes or heart disease

with a family history of breast cancer, adjusting for a BC PRS, was not significant (**Table 3.1**).

Genetic correlation provides insight into the shared genetic architecture of pairwise disease-family history

We examined the shared genetic architecture between all pairs of conditions and family history. The genetic correlation estimates for 63 pairwise combinations of disease and family histories of disease are shown in **Supplementary Figure 3.10**. We further examined the genetic correlation of disease with maternal, paternal, and sibling family history (**Supplementary Figure 3.11**). First, among significant ($P < 0.05$) same disease-family history pairs, we identified a strong positive genetic correlation ($r_g = 0.88$), where diabetes had the largest significant genetic correlation coefficient ($r_g = 0.98$), and prostate cancer had the smallest ($r_g = 0.75$).

Second, among non-same disease-family history pairs with a positive significant ($P < 0.05$) genetic correlation, we identified a moderately strong genetic correlation ($r_g = 0.31$), where the smallest and largest positive genetic correlation coefficients were diabetes and family history of stroke ($r_g = 0.13$) and depression and family history of prostate cancer ($r_g = 0.51$). Among non-same disease family history pairs with a negative (inverse) significant genetic correlation, we also identified a moderately strong negative genetic correlation ($r_g = -0.30$), where the smallest and largest negative genetic correlation coefficients were diabetes and family history of prostate cancer ($r_g = -0.13$) and emphysema/chronic bronchitis and family history of Parkinson's

Disease ($r_g = -0.50$). We identified an inverse genetic correlation between a family history of prostate cancer and several cardiometabolic traits, including heart disease (r_g = -0.30; p < 0.001), diabetes (r_g = -0.13), and stroke (r_g = -0.36), which is consistent with the protective effect we identified in our observational associations.

DISCUSSION

We present here an atlas of 132 non-same pairs of disease-family history associations for 12 complex human diseases and perform a range of analyses to assess the shared genetic architecture and environmental influences between a disease X and a family history of a different disease Y. Our findings are novel for several reasons. First, we demonstrate disease similarity for seemingly disparate disease conditions. Second, we identified novel associations, such as the protective effect of family history of prostate cancer and heart disease, a strong association between family history of emphysema/chronic bronchitis and depression, and an association between the prostate cancer condition and a family history of depression. These findings have not been reported to date and highlight a potential similarity between these two distinct phenotypes. Third, we extend our observational associations to investigate the effects of maternal and paternal history on cross-disease risk. For example, our findings suggest that maternal history of emphysema/chronic bronchitis exerts greater influence on depression than paternal history. Fourth, we disentangle genetic factors from environmental influences using a range of methods. For example, we find that the increased association between depression and family history of emphysema/chronic bronchitis in the adopted cohort compared to that of the non-adopted cohort may arise from shared environmental factors.

We also identified a greater magnitude of association in the adopted versus non-adopted cohort for depression and family history of stroke and emphysema and family history of lung cancer, supportive of stronger environmental influences contributing to risk in families with stroke and lung cancer.

We have also recapitulated several previously reported findings. For example, the positive associations between prostate cancer and family history of breast cancer as well as breast cancer and family history of prostate cancer is consistent with previous epidemiological findings [66,67]. We identified a strong association between family history of lung cancer and emphysema/chronic bronchitis, and additionally, with heart disease, high blood pressure, and diabetes, for which, we hypothesize is driven by smoking, a major risk factor for lung cancer, insulin resistance, and diabetes.

The findings of our study should be seen in light of several limitations. First, the same size of disease and family history for certain diseases are small (i.e., 137 individuals in the non-adopted cohort report Alzheimer's disease, only 4 individuals in the adopted cohort cohort report Alzheimer's disease, and 633 individuals in the adopted cohort report a family history of Alzheimer's disease), which could bias our observational associations and mean that our GWAS for family history is underpowered. Second, disease status and family history of disease were ascertained using self-reported questionnaires which are prone to measurement error and recall bias. Third, we assessed positive and negative family history by first-degree family history only (paternal, maternal, or sibling) and did not assess second or third-degree family history because this

information was not reported in the UK Biobank. Fourth, an adoption study makes an assumption that the adoptees are unrelated to their adoptive family.

XY-FamWAS broadens the search of family history as a risk factor for disease to 12 complex human diseases within a large cohort from the UK population. We demonstrate the clinical utility of family history for assessing disease risk, which can perform as well as PRS, but is also inexpensive and can be easily collected. We also perform a range of analyses that leverage genotypic information and an adoption study design for disentangling genetic effects from environmental influences of disease-family history associations, shedding light on biological mechanisms shared among seemingly different diseases. Our findings investigate shared genetic architecture and environmental influences among complex disease, pinpointing potential areas for early detection, prevention, or therapeutic intervention.

4
Mendelian Randomization Protocol

Conducting a Reproducible Mendelian Randomization Analysis using the R analytic toolkit [68]

Toolkit for performing a Mendelian randomization analysis in R using published summarized genetic data

Collaborator/co-author: Chirag J Patel
Manuscript published in *Current Protocols in Human Genetics* in January 2019

Significance Statement

Conventional observational epidemiological studies such as those presented in the previous chapters aimed at assessing the effect of an exposure on a disease phenotype can be subject to confounding such as reverse causation, where the disease precedes the

exposure[69]. A technique termed ‘Mendelian randomization’ (MR) can overcome this limitation by leveraging genetic variants such as single-nucleotide polymorphisms (SNPs) as instrumental variables to estimate exposure-outcome associations [70]. Summary statistics from genome-wide association studies (GWAS) facilitate conducting an MR analysis without the need for costly direct genotyping or obtaining individual-level data [71]. We describe here a protocol for assessing exposure-outcome associations in an MR framework using published GWAS summary statistics.

ABSTRACT

Mendelian randomization (MR) is defined as the utilization of genetic variants as instrumental variables to assess the causal relationship between an exposure and an outcome (Davey Smith & Ebrahim, 2003). By leveraging genetic polymorphisms as proxy for an exposure, the causal effect of an exposure on an outcome can be assessed while addressing susceptibility to biases prone to conventional observational studies, including confounding and reverse causation, where the outcome causes the exposure (Davey Smith & Ebrahim, 2007). Analogous to a randomized controlled trial where patients are randomly assigned to subgroups based on different treatments, in an MR analysis, the random allocation of alleles during meiosis from parent to offspring assigns individuals to different subgroups based on genetic variants (Davey Smith & Ebrahim, 2007). Recent methods use summary statistics from genome-wide association studies to perform MR, bypassing the need for individual-level data (Burgess et al., 2015). Here, we provide a straightforward protocol for using summary-level data to perform MR and provide guidance for utilizing available software.

INTRODUCTION

The aim of many, if not all, observational studies is to associate an exposure and a disease or phenotype to eventually collect evidence to discern a causal relationship. However, observational associations are influenced by biases such as measured and unmeasured confounding and reverse causality and therefore can lack ability to establish a directional effect[72]. The principle underlying Mendelian randomization (MR) methodology is that such biases can be circumvented by leveraging genetic variants associated with an exposure as an "instrumental variable" (IV) to estimate the effect of genetic variation within an exposure on an outcome[73]. An IV is defined as an external variable G that is associated with the exposure X and independent of outcome Y as well as any factors associated with outcome Y, other than via X [74]. Genetic variants can be utilized as "IVs", thereby serving the role of randomizing "exposure".

To utilize a genetic variant as an IV, three assumptions must be satisfied[75] (see Figure 4.1): (i) the genetic variant must be associated with the exposure, (ii) the genetic variant must be independent of any confounder of the exposure-outcome, and (iii) the genetic variant must be independent of the outcome, except via a possible association with the exposure.

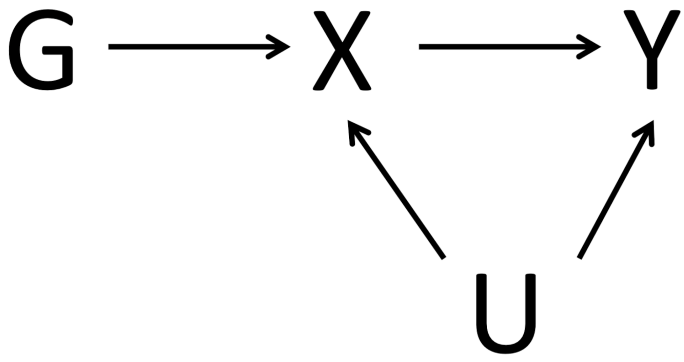


Figure 4.1. Directed acyclic graph depicting the IV assumptions for conducting Mendelian randomization. *G,* the genetic variant, must be (i) associated with exposure *X*, (ii) independent of any confounder *U*, and (iii) independent of outcome *Y*.

In the simplest MR technique (for one genetic variant), the presence of an association between a genetic variant and an exposure and the genetic variant and an outcome may imply causal effect of the exposure on the outcome[76]. MR can be performed with individual-level participant data, obtained from the genetic data for each participant, or with summary-level data, which usually contains per-allele regression coefficients and standard errors analyzed over all individuals within a study[77,78]. The causal effect of the exposure on the outcome can be calculated by a "2-stage least-squares" (2SLS) regression, where the exposure is regressed on the genetic instrument, and the outcome is regressed over the exposure values (where linear or logistic regression is used for continuous or binary outcome variables, respectively)[78]. In summary data MR, summary-level data can either be obtained from publicly available summary level data or by consortia of genome-wide association studies (GWAS), or can be calculated from individual-level participant information[71]. Here, we present a protocol to perform MR using summary-level data and we provide an RStudio markdown file to demonstrate how to use the TwoSampleMR package in R. The code and implementation of MR in the

protocols below are inspired by and utilize resources provided by the MRC Integrative Epidemiology Unit and the MR-Base Collaboration[79,80].

Performing a Mendelian randomization analysis in R using summarized genetic data

Introductory paragraph

In this protocol, we show how to perform MR using summary statistics, which can be applied to the one-sample or two-sample method. One-sample MR is performed when the data on the exposure and the outcome are derived from a single dataset[81]. Two-sample MR is performed when the data on the exposure and the outcome are derived from two non-overlapping and independent datasets, allowing one dataset to be used for performing the summary-level instrument-exposure analysis and the other dataset for performing the instrument-outcome association analysis[81,82].

In the inverse variance weighted (IVW) method, the causal effect of the exposure on the outcome for a single genetic variant can be estimated as a ratio of the association estimate for the outcome and the exposure [83,84]. For multiple independent genetic variants, the ratio estimates from each genetic variant can be meta-analyzed to form the overall causal estimate [83,84]. MR-Egger can be used when the IV assumptions do not hold or weakly hold, and entails a modification to the IVW estimate calculation where the intercept term is calculated as part of the MR-Egger estimate, instead of setting the intercept term of the regression to zero [85]. In MR-Egger, the intercept serves as a test for directional pleiotropy (meaning the genetic variants exert pleiotropic effects on the outcome)[83]. In the protocol below, we describe how to conduct an MR analysis using these methods and provide guidance for utilizing MR software in R in order to perform, interpret, and visualize results of MR analyses.

Protocol steps —*Step annotations*

1. Obtain GWAS summary statistics for your exposure (Figure 4.1, *X*) and outcome (Figure 4.1, *Y*) of interest. Resources such as the NHGRI-EBI Catalog[86] can be leveraged to search for and download publicly-available GWAS summary statistics.

2. Determine usability of GWAS summary statistics from Step 1 by ensuring that the instrument-exposure data and the instrument-outcome data have listed the effect allele, allele frequency, beta, standard error, p-value, and sample size (as shown in Figure 4.2).

CHR	POS	SNP	Tested_Allele	Other_Allele	Freq_Tested_Allele_in_HRS	BETA	SE	P	N
7	92383888	rs10	A	C	0.06431	0.0013	0.0042	0.7500	598895
12	126890980	rs1000000	A	G	0.22190	0.0001	0.0021	0.9600	689928
4	21618674	rs10000010	T	C	0.50860	-0.0001	0.0016	0.9400	785319
4	1357325	rs10000012	C	G	0.86340	0.0047	0.0025	0.0570	692463
4	37225069	rs10000013	A	C	0.77080	-0.0061	0.0021	0.0033	687856
4	84778125	rs10000017	T	C	0.22840	0.0041	0.0021	0.0480	686123

Figure 4.2. Shown are the first few rows of the body mass index GWAS summary statistics published from the UK Biobank and The Genetic Investigation of ANthropometric Traits (GIANT) Consortium meta-analysis[87].

3. Determine if the IV assumptions hold for conducting an MR analysis. The first assumption can be evaluated by linear regression of the exposure on the instrument and calculating the F-statistic for your instrument [88,89]. This can be calculated as, $F = \frac{N-K-1}{K} * \frac{R^2}{1-R^2}$, for N sample size, K number of genetic variants, and R^2 the proportion of the variance of the exposure explained by the IV [90]. An F statistic less than 10 denotes a weak instrument [88].

The second and third assumptions are more challenging to formally validate due to the possibility of unknown effects[88,89]. In assessing the second assumption, consider any potential confounding variables (Figure 4.1, U) that may play a role in the association between your exposure and outcome, and in assessing the third assumption, consider potential issues such as pleiotropy or population substructure that may serve as a violation [88,89].

4. Input exposure and outcome GWAS summary statistic data, using the read.table function.

```
exposure_data<-read.table("exposure_filename.txt", head=T, sep="\t") outcome_data<-
read.table("outcome_filename.txt", head=T, sep="\t")
```

5. Identify instruments. Find independent SNPs that are GWAS significant ($P < 5.0 \times 10^{-8}$) for the exposure and identify the effects for these instrument SNPs from the outcome GWAS.

6. Harmonize the exposure and outcome datasets. Ensure that the effect alleles from both files are the same. If not, then flip the log odds ratio of the effect allele of one of the datasets (multiply by -1). Ensure that the effect in the exposure file reflects the trait-increasing allele.

Ratio of coefficients (or Wald) method

7. Calculate the ratio of coefficients, or the Wald ratio. This is the simplest method for estimating the causal effect of the exposure on the outcome, and is the coefficient of the genetic variant in the regression of the outcome (represented here as outcome_data$beta) divided by the coefficient of the genetic variant in the regression of the exposure (represented here as exposure_data$beta)[91].

```
wald_ratio <- outcome_data$beta/exposure_data$beta
wald_ratio_standard_error <- outcome_data$SE/exposure_data$beta
z_statistic <- wald_ratio/wald_ratio_standard_error
p_value <- 2*pnorm(abs(z_statistic) ,lower.tail=F)
```

Note that the Wald ratio corresponds to the log odds ratio for the outcome per unit change of the exposure.

8. Perform a fixed-effects meta-analysis using the Wald ratio.

```
effect <- sum(wald_ratio*wald_ratio_standard_error^-2)/ (sum(wald_ratio_standard_error^-2))
standard_error <-  sqrt(1/sum(wald_ratio_standard_error^-2))
Z_statistic <- effect/standard_error
p_value <- 2*pnorm(abs(Z_statistic) ,lower.tail=F)
```

Inverse-variance weighted (IVW) method

9.Perform an inverse-variance weighted (IVW) linear regression to estimate the effect of the exposure on the outcome.

```
IVW_weights <- outcome_data$SE^-2
inverse_weighted_LR <- lm(outcome_data$beta ~ exposure_data$beta - 1
,weights=IVW_weights)
```

The command summary*(*inverse_weighted_LR) displays the effect, standard error, and p-value of the exposure on the outcome.

Note that the intercept term here is zero in order to calculate the IVW estimate [83]. In the case that a single genetic variant satisfies the IV assumptions, the effect of the exposure on the outcome can be estimated as a ratio of the estimated coefficient for the outcome to the estimated coefficient for the exposure for the genetic variant [83].

MR-Egger Regression

10. Perform an MR-Egger regression to estimate the effect of the exposure on the outcome.

```
MR_egger_regression <- lm(outcome_data$beta ~ exposure_data$beta,
weights=1/IVW_weights)
```

The command summary(MR_egger_regression) displays the effect, standard error, and p-value of the exposure on the outcome. Note that the intercept term here is calculated in the MR-Egger analysis [83,85].

Performing Mendelian randomization using the TwoSampleMR package in R.

The TwoSampleMR package in R facilitates conducting two-sample MR analyses by offering access to the large MR-Base repository of GWAS summary statistics and providing easy-to-use software for proper harmonization of datasets, estimating the causal effect using a range of MR methods, conducting sensitivity analyses, and visualizing results [79,80].

This protocol and code below was inspired by the TwoSampleMR documentation provided by the MRC Integrative Epidemiology Unit and the MR-Base Collaboration, which can be found on https://mrcieu.github.io/TwoSampleMR/[79,80].

Necessary Resources

Software

R package version >= 3.1.0 [42] with the following libraries installed: devtools[92], TwoSampleMR[79,80], MRInstruments[93],and tidyverse[94].

Files

GWAS summary statistics (including SNP, major allele, minor allele, allele frequency, effect size, standard error, p-value, and sample size) for the exposure and outcome of interest OR these files can be obtained by browsing through existing catalogues from the MR Base databases accessible through the MRInstruments package[93]. Note that some information that may be missing from your summary statistics file, may be present in the paper referencing the GWAS or may be calculated using the information in the file. Further note that your data can be formatted in the correct manner for use in the TwoSampleMR package by using the function format_data (as described in step #2 of the protocol below)[79,80].

The .Rmd file "TwoSampleMR_protocol.Rmd" included in this manuscript will serve as a guide through the protocol below.

Protocol steps—*Step annotations*

1. Load the TwoSampleMR package in R [79,80]. You can install the devtools package from CRAN-like repositories with the install.packages("devtools") command in order to utilize the install_github function[92].

```
install.packages("devtools")
library(devtools)
install_github("MRCIEU/TwoSampleMR")
library(TwoSampleMR)
```

2. Identify and obtain GWAS summary statistics. You can either obtain your own summary statistics or browse through the MR Base GWAS database[79] (available_outcomes() can show the list of available GWASs).

External summary statistics can be read in and converted to the correct format using format_data. For example, the body mass index (BMI) GWAS summary statistics as shown in Figure 4.2 can be converted as follows:

```
exposure_converted_dataframe <- format_data(exposure_dataset, type = "exposure", snp_col = "SNP", beta_col = "BETA", se_col = "SE", effect_allele_col = "Tested_Allele", other_allele_col = "Other_Allele", eaf_col = "Freq_Tested_Allele_in_HRS", pval_col = "P", samplesize_col = "N")
```

The R package MRInstruments contains data sources to search for genetic instruments that can be used for your MR analysis[93]. In this demonstration, we use data from the gwas_catalog to search for the instruments from the 2010 GWAS on BMI published in *Nature Genetics* by Speliotes et al [95].

```
devtools::install_github("MRCIEU/MRInstruments")
library(MRInstruments)
data(gwas_catalog)
exposure_data <- subset(gwas_catalog, PubmedID == "20935630")
```

3. Ensure that your data is presented in the correct input format and perform linkage disequilibrium (LD) clumping to remove any non-independent SNPs.

```
exposure_data <- format_data(exposure_data)
exposure_data <- clump_data(exposure_data)
```

4. Extract the instrumental SNPs for your outcome of interest. In this example, we are using the 2014 GWAS summary statistics for type 2 diabetes susceptibility as published in *Nature Genetics* by the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) consortium [96].

```
outcome_data <- extract_outcome_data(
   snps = exposure_data$SNP,
   outcomes = 23
)
```

5. Harmonize exposure and outcome datasets to ensure the reference alleles from both datasets match. Prune your harmonized dataset. Here, the exposure and outcome datasets are harmonized (shown in Figure 4.3) and renamed as dat.

```
dat <- harmonise_data(
   exposure_dat = exposure_data,
   outcome_dat = outcome_data
```

)

dat <- power.prune(dat)

```
> head(dat)
        SNP effect_allele.exposure other_allele.exposure effect_allele.outcome other_allele.outcome beta.exposure beta.outcome
1  rs2444217                     A                    G                     A                    G            NA -0.009950331
2   rs255414                     A                    G                     A                    G            NA -0.029558802
3  rs2922763                     T                    G                     T                    G            NA  0.009950331
4  rs3764400                     C                    T                     C                    T            NA -0.009950331
5   rs867559                     G                    A                     G                    A            NA  0.019802627
6 rs10150332                     C                    T                     C                    T          0.13  0.048790164
```

Figure 4.3. Shown are the first few rows of the harmonized dataset.

6. Perform an MR analysis (results shown in Figure 4.4) and specify the types of method in method_list() of the mr()function.

results <- mr(dat)

```
  id.exposure id.outcome                outcome                       exposure                    method nsnp         b         se         pval
1      zQIvwv         23 Type 2 diabetes || id:23 Body mass index (kg/m2 increase)                MR Egger   24 0.2463348 0.06362685 8.245661e-04
2      zQIvwv         23 Type 2 diabetes || id:23 Body mass index (kg/m2 increase)          Weighted median   24 0.1809661 0.03865444 2.845919e-06
3      zQIvwv         23 Type 2 diabetes || id:23 Body mass index (kg/m2 increase) Inverse variance weighted   24 0.1908652 0.02836724 1.715750e-11
4      zQIvwv         23 Type 2 diabetes || id:23 Body mass index (kg/m2 increase)              Simple mode   24 0.2248712 0.06910506 3.495706e-03
5      zQIvwv         23 Type 2 diabetes || id:23 Body mass index (kg/m2 increase)            Weighted mode   24 0.1979986 0.04933998 5.447451e-04
```

Figure 4.4. The causal effects, standard errors, and p-values obtained from the MR analysis using the default methods of MR Egger, weighted median, inverse variance weighted, simple mode, and weighted mode, are shown.

The full list of available MR methods can be identified from mr_method_list().

7. Conduct sensitivity analyses. Check for heterogeneity and test for directional horizontal pleiotropy.

mr_heterogeneity(dat)
mr_pleiotropy_test(dat)

8. Perform a leave-one-out sensitivity analysis (by sequentially removing each SNP from the MR analysis and running MR) and visualize results from this sensitivity analysis (shown in Figure 4.5).

results_leaveoneout <- mr_leaveoneout(dat).
mr_leaveoneout_plot(results_leaveoneout)
plot_leaveonout[[1]]

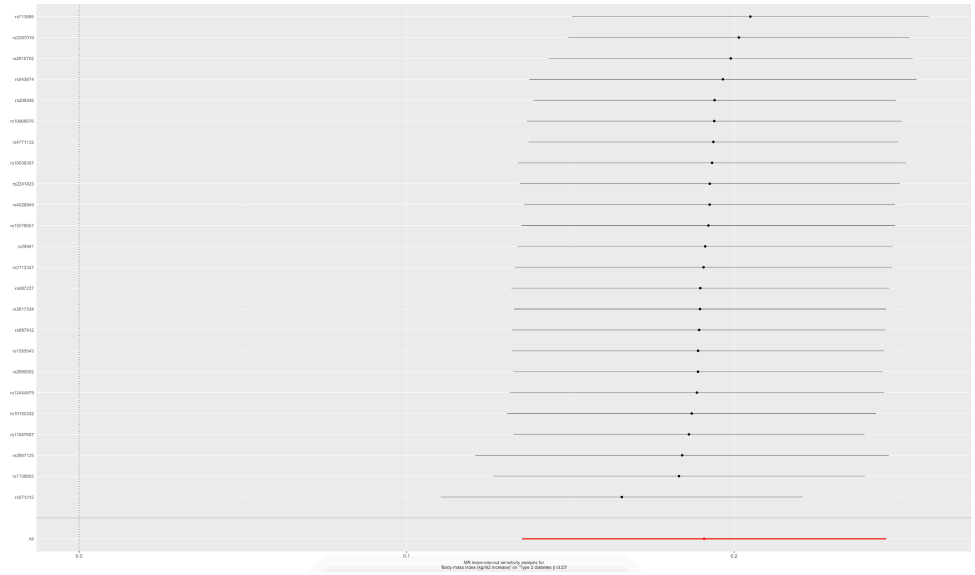


Figure 4.5. The results from the leave-one-out sensitivity analyses are shown on the scatterplot. The estimated causal effect is shown for each excluded SNP and the overall estimate using all the SNPs is shown in red. The error bars represent the 95% confidence intervals.

9. Visualize MR results.

```
scatter_plot <- mr_scatter_plot(results, dat)
scatter_plot[[1]]
```

The command mr_scatter_plot(results, dat) creates a scatterplot for each exposure-outcome association (shown in Figure 4.6). A specification of the method in method_list() visualizes the estimated causal effect according to the specified MR method.

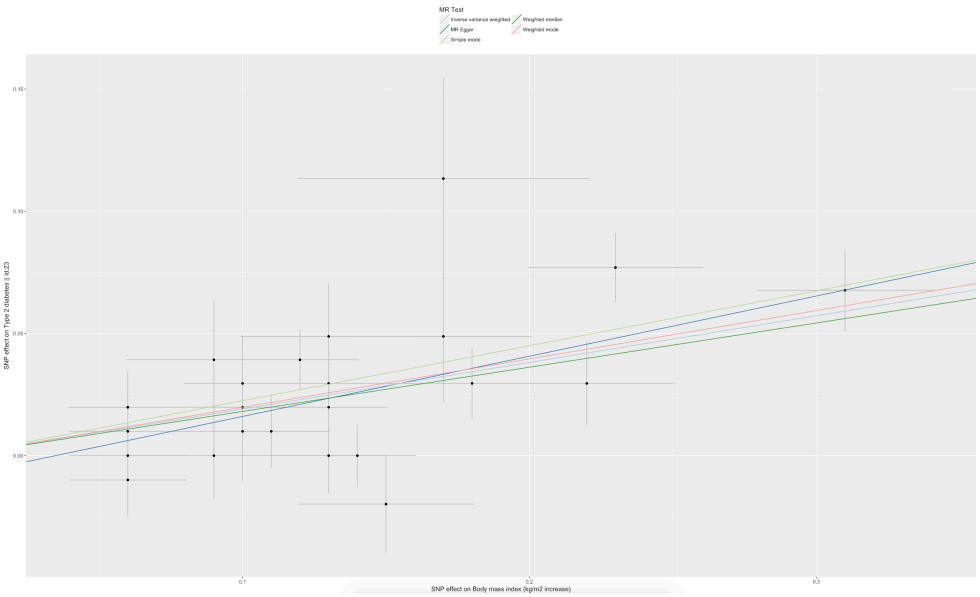


Figure 4.6. The scatterplot suggests a positive causal relationship of the SNP effects on BMI against the SNP effects on type 2 diabetes. Each point displayed on the graph represents a single genetic variant. The horizontal and vertical lines extending from each point represent the 95% confidence interval for the genetic associations. The horizontal axis of the graph displays the estimated genetic associations with the exposure (BMI), and the vertical axis displays the estimated genetic associations with the outcome (type 2 diabetes). The color of the lines indicate the type of MR test used (light blue for inverse variance weighted, dark blue for MR Egger, light green for simple mode, dark green for weighted median, and red for weighted mode).

Additionally, a forest plot can be made to compare the MR estimates derived from the different MR methods (shown in Figure 4.7).

```
single_snp_analysis <- mr_singlesnp(dat)
forest_plot <- mr_forest_plot(single_snp_analysis)
forest_plot[1]
```

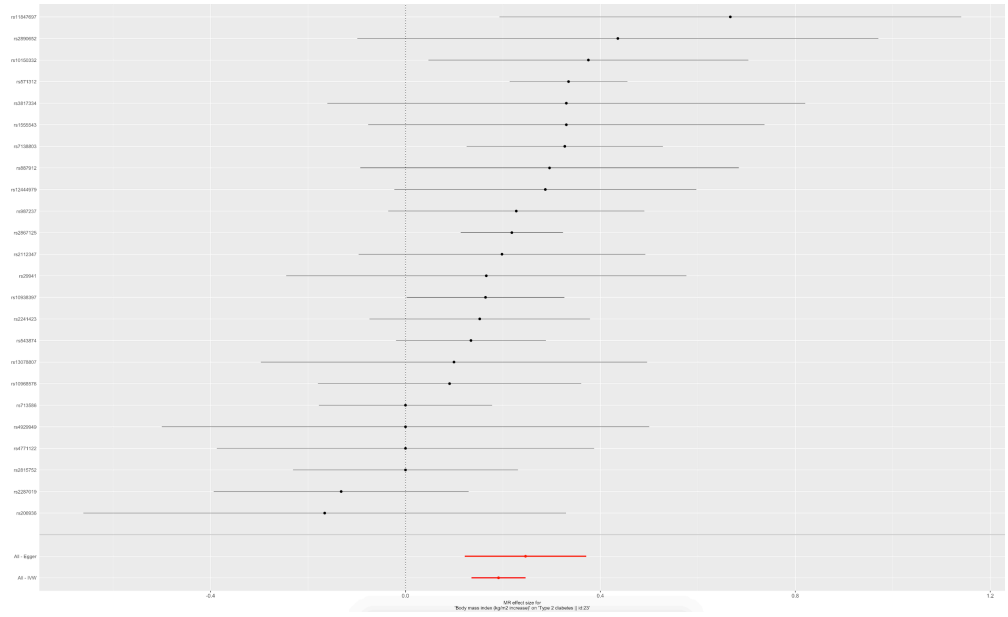


Figure 4.7. The forest plot shows the causal estimate using each SNP alone as well as the overall causal estimate using all the SNPs with MR-Egger and IVW. The error bars represent the 95% confidence intervals.

GUIDELINES FOR UNDERSTANDING RESULTS

By leveraging a genetic approach as demonstrated in our example above, we were able to provide evidence in support of a positive causal effect of BMI on type 2 diabetes, which

was consistent across all MR methods. We obtained effect sizes of 0.25, 0.18, and 0.19 for MR Egger, weighted median, and inverse variance weighted, respectively, which correspond to the estimated causal effect on type 2 diabetes per unit increase in BMI (kg/m^2). In a "leave-one-out" sensitivity analysis, where we sequentially excluded a SNP and performed MR, we observe that the causal estimate remains robust. The forest plot compares the estimated causal effects for all the SNPs as determined by MR-Egger and IVW to the estimated causal effect as determined per each SNP. While the MR-Egger and IVW estimates agree in our demonstrated example, the IVW estimate can substantially differ from the MR-Egger estimate, suggesting the possibility of directional pleiotropy [83]. In summary, we highlight the utility of MR in assessing causal relationships, while accounting for limitations prone to many conventional observational epidemiological studies.

COMMENTARY

Background Information

The concept of utilizing IVs to examine causal effects was first introduced in econometrics 90 years ago, and applied to disease outcomes in 1986 by Martijn Katan [97]. In assessing the causal role of low serum cholesterol levels and cancer, Katan explained that the relationship was likely not affected by diet or other confounding factor, but that the relationship can be elucidated by observation of the number of cancer patients who carry the E-2 isoform of the apolipoprotein (ApoE) gene, which is associated with lower serum density lipoprotein than major isoforms E-3 and E-4 [98]. Since then, there have been many studies that have assessed causal relationships using MR for a range of exposures and outcomes, including biomarkers (i.e. C reactive protein[99]), clinical traits

(i.e. BMI [100]), disease phenotypes (i.e. coronary heart disease [101]), socioeconomics (i.e. educational attainment [102]), behavioral characteristics (i.e., alcohol consumption [103]), and intrauterine effects[104] (i.e., maternal homocysteine levels[105]).

For example, an MR study demonstrated that genetic variants in the gene encoding the target of statin therapy, HMG-CoA reductase or *HMGCR*, is associated with increased risk for type 2 diabetes and related traits such as higher body weight and waist circumference, highlighting a pharmacological application of MR[106]. In another example, MR was used to determine that tobacco smoking may cause a reduced BMI and a higher resting heart rate, but did not find a strong causal association between smoking and adverse blood pressure, serum lipids, and glucose levels[107]. MR promises to be a valuable method for identifying disease risk factors and areas for intervention and can be leveraged to inform public health policy.

Critical Parameters

There are a number of statistical and methodological challenges and limitations to MR that have been discussed at length in other articles [78,108,109]. Possible limitations include linkage disequilibrium (i.e., when different loci within a population have correlated allelic states[76]), population stratification (i.e., when a population can be broken into subpopulations that exhibit different frequencies of genetic variants or disease[76]), or pleiotropy (i.e., when a genetic variant is associated with more than one phenotype[76]). Challenges may arise from utilizing a weak instrument (F statistic less than 10), or from situations where the core assumptions are violated or weakly satisfied,

and even from cases where the core assumptions are satisfied, but an external factor is at play (i.e., canalization) [110]. In fact, the development of novel MR approaches and extensions to the conventional methodology to account for these limitations is a rapidly growing field [111–115].

For a description of potential limitations that may affect interpretation of MR findings and recommended practices in those situations, we recommend referring to *Table 2* from a review article by Zheng [110] and *Table II* from Lawlor [76]. We also recommend referring to *Table 2* from the review article by Burgess for descriptions of various sensitivity analyses and situations where they would be of relevance [116].

ACKNOWLEDGEMENT

We thank Dr. George Davey Smith and his team at the MRC Integrative Epidemiology Unit at the University of Bristol for offering the short course at the 2017 Mendelian Randomization Conference and for providing resources for conducting an MR study. We also thank the MR-Base Collaboration for providing the extended documentation for the *TwoSampleMR* package (accessible at: https://mrcieu.github.io/TwoSampleMR/).

5
Infection-Wide Association Study (IWAS) for Type 2 Diabetes

Magnitude and robustness over time of infectious risks in patients with documented type 2 diabetes: an infection-wide association study ("IWAS") and Mendelian randomization assessment

Collaborators/co-authors: Arjun K Manrai, John PA Ioannidis, Chirag J Patel

ABSTRACT

Background. Some infectious diseases are speculated to predispose individuals to develop type 2 diabetes (T2D) and T2D increases the risk of many infections. However, these studies lack comprehensive testing of the full spectrum of all infectious disease diagnoses.

Methods. We performed an observational investigation, "Infection-Wide Association Study" (IWAS), examining 252 and 274 diagnostic codes for infectious diseases recorded before ("IWAS-b") and after ("IWAS-a") the first date of documented T2D in 172,172 individuals with T2D versus matched controls from 44·9M members from large health insurance claims. We provide a comprehensive picture of the magnitude of infectious disease associations with T2D over six years surrounding the time of T2D diagnosis and link infectious disease susceptibility genotypes with T2D using Mendelian randomization (MR).

Findings. We identified 31 (12% of total) and 28 (10%) diagnostic codes associated with T2D in IWAS-b and IWAS-a, respectively, of which 23 infections were identified in both. In IWAS-b, we identified increased odds of candidiasis, Hepatitis C, *Staphylococcus aureus*, *Streptococcus*, and *Helicobacter pylori*, which increased in magnitude of odds following T2D diagnosis. The MR analyses showed no significant signals, except for a suggestive signal for Hepatitis B (P = 0.008).

Interpretation. In patients with T2D, many infections show early signals of increased risk and the risk escalates substantially once the diagnosis of T2D is formally recorded. These infections create a high burden of complications but even though they can occur early in the T2D disease process they do not appear etiologically related as risk factors with T2D.

Research in context

Evidence before this study: We searched PubMed for articles that have assessed the relationship of infectious diseases with type 2 diabetes published up to October 15, 2019 using a combination of search terms that included synonyms of "infectious diseases", "type 2 diabetes", "bacterial infections", and "viral infections". Our search identified reports from observational epidemiological studies where one or a few infectious diseases were studied in association with type 2 diabetes. These studies lacked comprehensive assessment of the full spectrum of all infectious disease diagnoses studied as potential risk factors and complications of type 2 diabetes. Further, evidence from conventional observational studies cannot establish direction of association.

Added value of this study: In our observational investigation, termed "Infection-Wide Association Study" (IWAS), we comprehensively and systematically assessed over 250 diagnostic codes for infectious diseases recorded before and after the first date of documented type 2 diabetes (T2D) in over 170,000 patients and controls in a cohort assembled from health insurance claims data. We catalogue a comprehensive list of infections, including viral, bacterial, fungal, and parasitic diseases, presented at point-of-care in T2D patients over a period of three years before and after the date of documented T2D and provide a granular picture of the magnitude of associations of infectious diseases with T2D over one-year time increments during this six-year period. Our method allows for discovery of infectious diseases that may be possible risk factors of T2D and/or complications. We integrated our analyses with the Mendelian randomization approach to assess causality in associations observed from IWAS, enhancing our

understanding of the bidirectional link between T2D and infectious diseases and allowing us to prioritize infectious diseases associated with T2D.

Implications of all the available evidence: Our study presents a data-driven map of infectious diseases evaluated in association with T2D as risk factors and/or complications across the broad scale presented in large claims, and provides a genetic assessment of the causal relationship of prioritized infectious diseases and T2D. Our findings show individuals with T2D are at risk for a high burden of infectious disease complications that occur early in the T2D disease process, but do not appear to be etiologically related as risk factors with T2D.

INTRODUCTION

The pathogenesis of type 2 diabetes mellitus (T2D), a metabolic disorder characterized by β-cell dysfunction in insulin secretion, involves many components[117]. Heritability estimates for T2D derived from family and twin studies have varied from 20-80%, indicating a potential contribution in the development of abnormal glucose tolerance explained by environmental factors[12,118]. It is unclear to what extent infectious agents could play a role.

Previous studies have implicated a number of infections that can potentially play a role in T2D, but disentangling the mechanisms for the abrupt metabolic change remains at large [119,120]. These studies are limited to addressing one type of infection or infectious agent at a time and lack comprehensive assessment of the full spectrum of infectious

agents. Assessing these factors in a non-systematic and non-standardized fashion can lead to spurious findings and a fragmented literature of associations[121–123]. Furthermore, the role of infectious agents may be complex in T2D. Some infections may act as a risk factor for insulin resistance or beta-cell function before disease onset.

While the role of infectious risk factors is controversial, individuals with T2D may have risk for infection as complications due to high glucose levels. There is literature on assessing the risks of each one of them one at a time[124,125], but we lack a comprehensive analysis to-date that examines comparatively the magnitude of the risk for all infections after T2D has been diagnosed. Some infections may also occur both before and after T2D onset. The exact timing of the T2D process may be difficult to discern and may cause difficulties to specify whether an infection preceded or followed the onset of that process.

To enhance our understanding of the associations of infectious disease factors both before T2D documentation, we performed an "Infection-Wide Association Study" (IWAS) in cohorts assembled from health insurance claims data. We comprehensively and systematically evaluate the associations of 252 and 274 infectious diseases that occurred within a 24-month window before ("IWAS-b) and after ("IWAS-a") the date of documented T2D in a study population of 172,172 patients with T2D and matched controls. Our approach is conceptually similar to that of genome-wide association studies (GWAS) and exposure-wide associations (EWAS)[126]. Specifically, we take an agnostic approach and scan across the broad spectrum of infectious diseases for

association with T2D while correcting for multiple testing. We further investigated associations at varying time intervals within a three-year period before and after the documented date of diagnosis to probe the robustness of the results to different time-of-disease onset assumptions.

Observational investigations can be prone to biases including measured and unmeasured confounders and T2D onset may precede the recording of a T2D diagnosis in medical care-related databases. Therefore, we coupled our findings with Mendelian randomization (MR) analyses in order to assess the potentially causal effect of our top findings (lowest false discovery rate [FDR]) on T2D. Our study is the first to combine IWAS and MR to provide a comprehensive picture of the magnitude of increased infectious risks before and after a documented diagnosis of T2D, and to examine whether infections are not only complications but even risk factors per se for T2D.

METHODS

Study population

We used de-identified medical and pharmacy claims data from Aetna Inc., a large national insurance company in the United States. The retrospective dataset contained 44·9 million members with billed medical services from January 2008 to February 2016 across 50 states and territories. Information about the members of the insurance plan included member's age, sex, enrollment period, as well as diagnoses, prescription history, and laboratory test results collected in the medical billing claims processes. For each

member included in our study, we searched for T2D billing codes and we identified infectious disease claims during a 24-month surveillance window before and the documented date of T2D diagnosis and a 24-month surveillance window after that date (**Figure 5.1A, Continued**). We selected a 24-month surveillance window because glycemic measurements appear to rapidly change 2-3 years prior to the onset of diabetes[127].

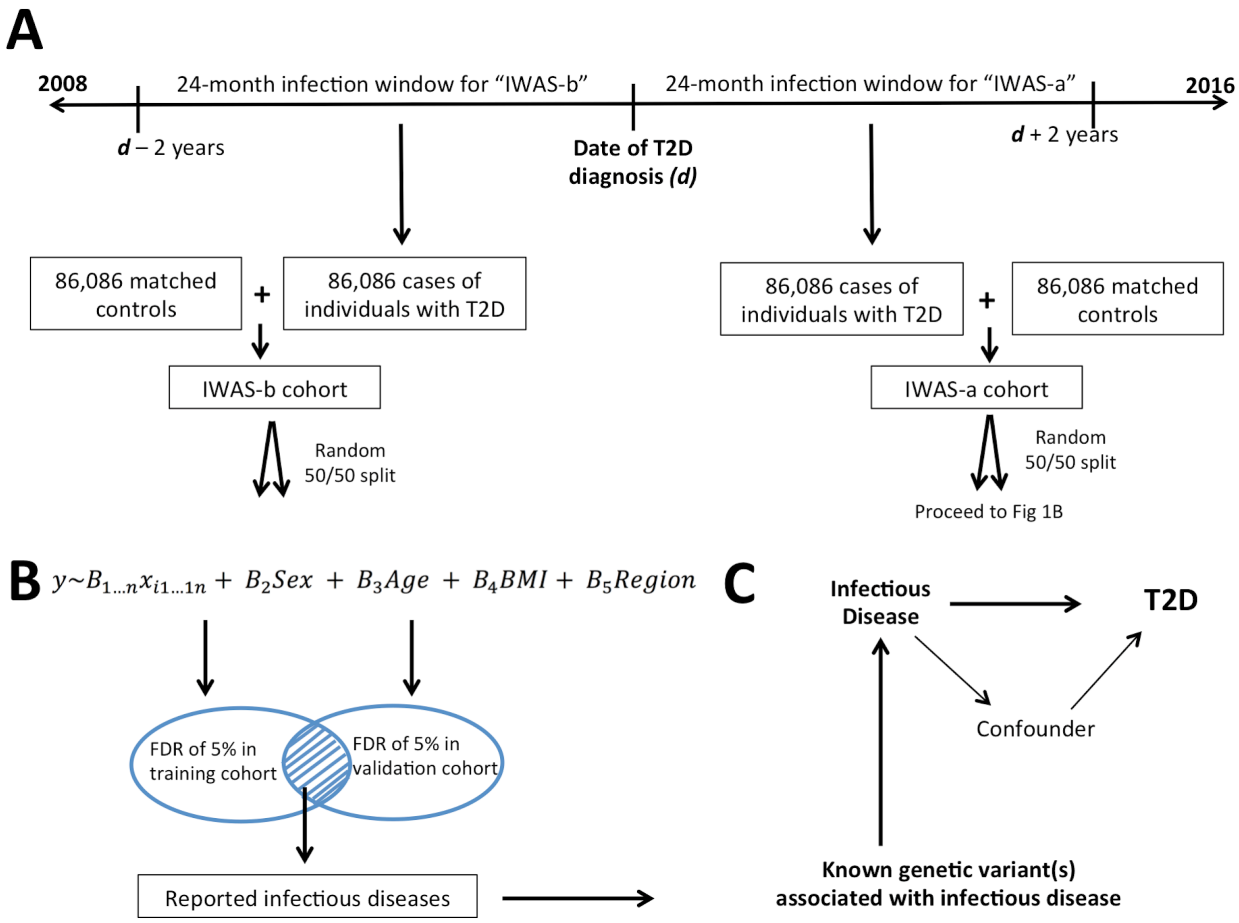


Figure 5.1. Diagram of IWAS case-control ascertainment and method of analysis. A) We ascertained our cases by identifying patients with an ICD-9 diagnostic billing code for type 2 diabetes (T2D), where the "date of T2D diagnosis" was defined as the earliest date this code was entered such that patients did not have a prescribed T2D medication or blood glucose or hemoglobin A1C test result indicative of T2D prior to this documented date. We identified 86,086 cases that met our selection criteria and 86,086 propensity score matched controls based on age, gender, census region, and overweight/obesity. We identified all infectious disease ICD-9 diagnostic codes that the cases had in the 24-month window before and after the date of T2D diagnosis for our IWAS-b and IWAS-a analyses, respectively, and that the controls had within the same 24-month window of surveillance as the matched cases. **B)** We randomly allocated half of the cases and half of the controls to a training and validation set. We then systematically tested 252 and 274 infectious disease diagnostic codes for their association with T2D for IWAS-b and IWAS-a, respectively. We report infectious disease diagnostic codes that met an FDR significance threshold of 5% in both the training and validation sets for IWAS-b, and separately for IWAS-a. **C)** We examined the potential causal association of each identified infectious disease from IWAS-b with T2D in a Mendelian randomization framework.

Ascertainment of T2D diagnosis

We ascertained documented diabetic status by the presence of an International Classification of Diseases, Ninth Revision (ICD-9) diagnostic billing code in the range 250-250·92 during any hospital visit including outpatient and inpatient **(Supplementary Table 5.1)**. We excluded codes in the range that were specified for type 1 diabetes and removed patients under age 18 to further ensure exclusion of type 1 diabetes cases. For every patient, we determined the date of T2D diagnosis by identifying the earliest documented date a patient received an ICD-9 code for T2D (referred here as "date of T2D diagnosis"). We excluded patients with a minimum enrollment period less than 24 months prior to the date of T2D diagnosis, resulting in a sample population of 296,497 patients **(Supplementary Figure 5.1)**. We further excluded 145,628 patients without any prescription history, which can be suggestive of incomplete records, and another 17,912 patients with a blood glucose or hemoglobin A1C test result prior to the date of T2D diagnosis that may be indicative of T2D in accordance with guidelines set by the American Diabetes Association[33]. Specifically, this included: (1) fasting glucose level greater than 7·00 mmol/L (126 mg/dL); (2) non-fasting glucose level greater than 11·1 mmol/L (200 mg/dL); and (3) hemoglobin A1C concentration greater than 6·5%[33]. We identified laboratory test results by entries for LOINC (Logical Observation Identifiers Names and Codes) 1558-6, 2345-7, and 4548-4 for fasting glucose, non-fasting glucose, and hemoglobin A1C, respectively. To ensure we determined the date of T2D diagnosis and that the patients did not have T2D prior to enrolling in the insurance plan, we excluded 13,725 patients who were prescribed a T2D medication prior to the recorded

date of T2D diagnosis (**Supplementary Table 5.2**). We identified medications prescribed during the study period using the National Drug Code (NDC) description[128].

For every patient, we identified age at time of T2D diagnosis, gender, and census region (Northeast, Midwest, South, and West). We also identified body mass index for each patient by searching for an ICD-9 code (see **Supplementary Table 5.3** for list of codes) within two years prior to the date of T2D diagnosis. We excluded 33,146 patients with inconsistent patient identification numbers or demographic information, resulting in a final case population of 86,086 patients **(Figure 5.2)**.

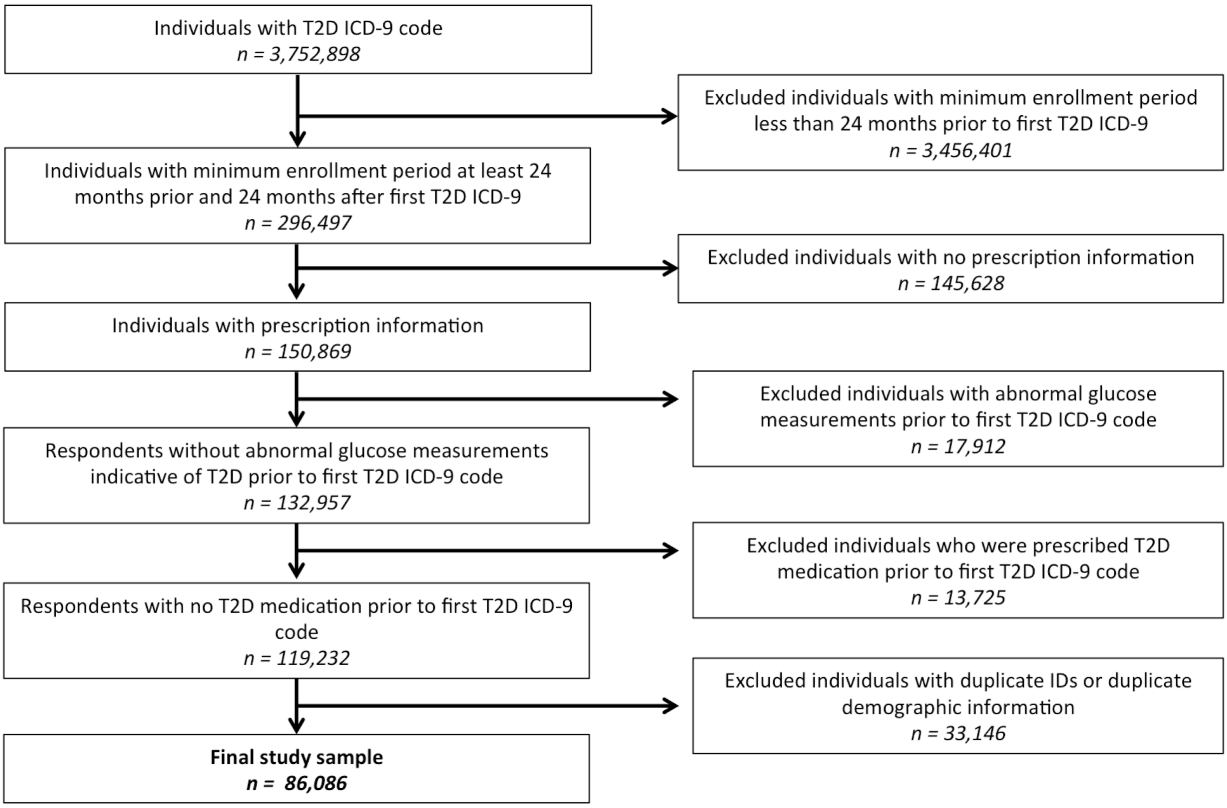


Figure 5.2. Flowchart depicting inclusion and exclusion criteria for study case selection.

Matched controls by propensity score

We randomly selected 410,000 individuals who did not have any ICD-9 billing code for T2D and who were not prescribed any T2D medications or insulin therapy nor have had a blood glucose or hemoglobin A1C laboratory test with a result indicative of T2D at any point in the insurance claims dataset. Consistent with our ascertainment of cases, we ensured patients were not without a prescription history, and were over age 18. For each control, we identified date of birth, gender, census region, and body mass index.

We implemented "nearest-neighbor" propensity score matching to match controls in a 1-to-1 ratio to cases based on age, gender, census region, and overweight/obesity status. Consistent with cases, we defined overweight/obesity status in controls by the presence of a corresponding body mass index billing code within two years prior to the age of matching. We clustered diabetic cases by year of T2D billing code, which ranged from 2010 to 2015, and for each cluster, we performed greedy nearest neighbor matching without replacement. For every selected case, we matched a control subject based on distance of their logistic regression-derived propensity score. We then identified infectious disease ICD-9 codes that occurred within the same 24-month infection window as their matched cases, which was a 24-month window before the date of matching for IWAS-b, and 24-months after the date of matching for IWAS-a. For the controls, we calculated their age at the year the control was binned in (i.e., if a control was matched to a case whose date of T2D was in 2010, then the age of the control was taken in 2010, in accordance with the age of the matched case). All propensity score matching was performed using the program *MatchIt* in the R statistical computing environment[129].

Systematic identification of infections before and after type 2 diabetes

We performed an observational investigation, termed "Infection-Wide Association Study" (*IWAS*), on the case and control groups to identify infectious diseases that potentially increase an individual's likelihood of developing T2D (termed "IWAS for infections **b**efore the date of T2D" or IWAS-b) or that would follow as complications of T2D (termed "IWAS for infection complications **a**fter the date of T2D" or IWAS-a). Following the construction of the case and control groups, we extracted patient-level ICD-9 codes for infectious diseases discerned by an ICD-9 code in the range of 0 to 139 during the 24-month period prior to the date of T2D for our IWAS-b analyses. We additionally conducted this analysis using the 24-month period after the date of T2D as our infection surveillance window for our IWAS-a analyses. We removed 130 and 109 infections from IWAS-b and IWAS-a, respectively, that occurred in fewer than 5 patients (or with a prevalence less than 0·00006%) of the case and control groups combined. We examined 252 and 274 unique codes for infectious disease in IWAS-b and IWAS-a, respectively, where 225 unique codes overlapped between IWAS-b and IWAS-a. The difference between the number of infections examined in IWAS-b and IWAS-a is largely due to the prevalence inclusion threshold. We studied infectious diseases across a wide range of categories, with 75% of the examined infectious diseases represented as viral, bacterial, fungal, and parasitic diseases (displayed in **Supplementary Table 5.4**).

For every patient in the study, we marked with a binary variable (0 or 1) whether the patient had a billing code for an infectious disease at any point during the 24-month

window of surveillance. We did not take repeated incidences of infectious disease codes for the same patient into account because this could occur due to billing processes across clinics and may not necessarily represent cases of unique infection. Of the patients in the study, half were randomly assigned for infectious disease identification, and the other half was used for internal replication. We modeled the relationship between each infection and T2D using logistic regression, adjusting by age, sex, census region, and body mass index as a continuous variable **(Figure 5.1B)**. We corrected regression coefficient significance levels for multiple testing using the FDR method[130].

Time-Varying IWAS Analyses

The documented date of new-onset T2D is a terminus ante quem for when T2D occurred. The diabetic process or even frank T2D may have started earlier than that date but not be documented because no testing was done promptly, because testing cannot identify early phases of the diabetic process, or because it is not captured in the patient's insurance claims record. In order to gain a more granular time-varying view beyond our primary 24-month pre- and post- date of documented T2D analyses, we extended our study to evaluate associations at varying time intervals within a three year period before and after the date of documented T2D. The varying time windows examine the stability of associations with different temporality assumptions for IWAS-b. Similarly, the varying time windows allow assessing time variability after the documented onset of T2D in the IWAS-a associations and their strength.

More specifically, we identified infectious disease ICD-9 billing codes that occurred within 0-12, 12-24, and 24-36 months before and after the date of documented T2D. If a patient had multiple incidences of an infectious disease (i.e., an infection 18 months prior to the date of T2D, and again 6 months prior), then we marked the patient as having the disease within multiple surveillance windows (i.e., both the "0-12" and "12-24" month surveillance windows). We then systematically tested each infection for association with T2D for each of the three time intervals before and after the date of documented T2D using logistic regression, adjusting by age, sex, census region, and body mass index.

Mendelian randomization analyses

In order to evaluate the potential causal effect of infection on T2D, we performed two-sample Mendelian randomization (MR) on infectious diseases that met an FDR threshold of 5% in the training and validation sets of IWAS-b **(Figure 5.1C)**. We identified genetic variants associated with the infectious phenotype of interest that met a genome-wide significance threshold in published genome-wide association studies (GWAS) by searching for publicly available summary statistics in the NHGRI GWAS Catalog[131] and the MR-Base repository[79,132] and used these genetic variants as proxy for the infectious exposure.

We examined T2D, hemoglobin A1C, levels, and fasting insulin levels as outcomes in our MR analyses. For T2D, we used data from the DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) consortium, which was a meta-analysis on 26,488 cases and 83,964 controls of European, East Asian, South Asian, Mexican, and Mexican American

ancestry[96]. For hemoglobin A1C levels, we used data from a meta-analysis of 23 GWAS in 46,368 nondiabetic individuals of European ancestry, and for fasting insulin, we used data from a meta-analysis of 21 GWAS in 46,186 nondiabetic participants from the Meta-Analyses of Glucose and Insulin-Related Traits Consortium (MAGIC)[133,134].

For *Helicobacter pylori*, we obtained summary statistics from a GWAS on seroprevalence in 6160 individuals of European descent who were seropositive for *H. pylori[135]*. We identified 2 instrument SNPs, rs10004195 (P = 1·4e-18) and rs368433 (P = 2·1e-8), and that were not in linkage disequilibrium[135]. We identified two genetic instruments, rs225126 (P = 3·0e-10) and rs7161578 (P = 4·0e-8), for yeast infection from a GWAS study with data from 52,218 research participants of European ancestry from 23andMe[136]. For human immunodeficiency virus (HIV), we leveraged one genetic instrument, rs4878712, associated with HIV-1 susceptibility[137]. For hepatitis B, we used a GWAS on 1425 cases of European descent with hepatitis B and 218,180 European ancestry controls, where an association was found for rs9268652 (P = 3·1e-9)[136]. For hepatitis C, we used one genetic instrument, rs8099917 (P = 6·0e-9), associated with chronic hepatitis C in 1015 cases of European descent[138].

We also performed a two-sample MR on antibody level in response to infection and T2D. We identified a genome-wide significant genetic variant for antibody level in response to infection from a GWAS study on 1300 Mexican Americans who were measured for IgG antibody level against 12 common infections, including *Chlamydia pneumoniae, H.*

pylori, Toxoplasma gondii; cytomegalovirus; herpes simplex I virus; herpes simplex II virus; human herpesvirus 6 (HHV6); human herpesvirus 8 (HHV8); varicella zoster virus; hepatitis A virus (HAV); influenza A virus; and influenza B virus[139].

We performed all two-sample Mendelian randomization analyses using the TwoSampleMR package in R[79,80] and derived estimates using the Wald ratio and inverse-variance weighted methods.

RESULTS

Characteristics of study population

Table 5.1 (continued) shows the demographic characteristics for the IWAS-b and IWAS-a study populations, respectively, displayed separately for the training and testing sets, as well as for the cases and controls. There were 172,172 total patients in IWAS-b and IWAS-a study populations, respectively. The average age across both of these populations was 54 and 50% were female, with 18% of the population classified as overweight or obese **(Table 5.1)**. Of the patients with ICD-9 diagnostic codes for body mass index (BMI), the average BMI was 32. Patients were distributed across the four Census Bureau-designated regions, with 32% from the Northeast, 12% from the Midwest, 38% from the South, and 18% from the West.

Table 5.1. Demographic characteristics of case and control groups in IWAS-b and IWAS-a.

	IWAS-b						IWAS-a					
	Training			Validation			Training			Validation		
	All	Has T2D	Does not have T2D	All	Has T2D	Does not have T2D	All	Has T2D	Does not have T2D	All	Has T2D	Does not have T2D
No. of patients	86,086	43,043	43,043	86,086	43,043	43,043	86,086	43,043	43,043	86,086	43,043	43,043
Age, mean	53.9	53.8	54.0	54.0	54.0	54.0	53.9	53.8	54.0	53.9	54.0	53.9
Female (%)	49.9	49.7	50.1	49.9	49.9	49.8	49.7	49.5	49.8	50.0	50.1	50.0
Region (%)												
West	17.8	17.8	17.8	17.9	17.7	18.0	17.9	18.0	17.8	17.8	17.4	18.1
Midwest	12.0	11.8	12.2	11.8	11.7	11.9	11.8	11.8	11.8	11.9	11.8	11.9
Northeast	32.3	32.7	32.0	32.6	32.6	32.5	32.6	32.4	32.9	32.5	33.0	32.1
South	37.9	37.7	38.0	37.8	38.0	37.5	37.7	37.8	37.5	37.8	37.8	37.8
Overweight or obese (%)	17.8	17.8	18.2	18.0	17.9	18.1	18.0	17.8	18.2	18.0	17.9	18.1
BMI, mean	32.0	32.8	31.3	32.0	32.7	31.2	32.0	32.7	31.3	32.0	32.8	31.3

We identified 252 and 274 eligible ICD-9 diagnostic codes from the IWAS-b and IWAS-a study populations, respectively, that represented a range of infectious diseases **(Supplementary Table 5.4)**. Over 70% of the ICD-9 diagnostic codes represented viral, bacterial, fungal, and parasitic diseases.

Infection associations with T2D before documented date of T2D

Figure 5.2 (continued) displays a "Manhattan plot" style figure for findings from the infection-wide association study of infections that occurred before the documented date of T2D diagnosis (IWAS-b). The negative logarithm of the association P-value is displayed on the y-axis, where the height corresponds to the strength of the association to T2D, displayed for each infection ICD-9 code (displayed along the x-axis). We identified

31 of 252 (12·3%) infection ICD-9 diagnostic codes associated with T2D, at an FDR threshold of 5% in both the training and validation sets **(Figure 5.2, continued)**. The most significant (lowest FDR) findings included: dermatophytosis of foot (OR 4·11, P = 2·4E-21) and nail (OR 3·58, P = 5·3e-22), unspecified septicemia (OR ·4e-21), viral warts (OR 2·41, P = 9·5E-18), and *H. pylori* (OR 3·03, P = 6·1E-13) **(Supplementary Figure 5.1a)**.

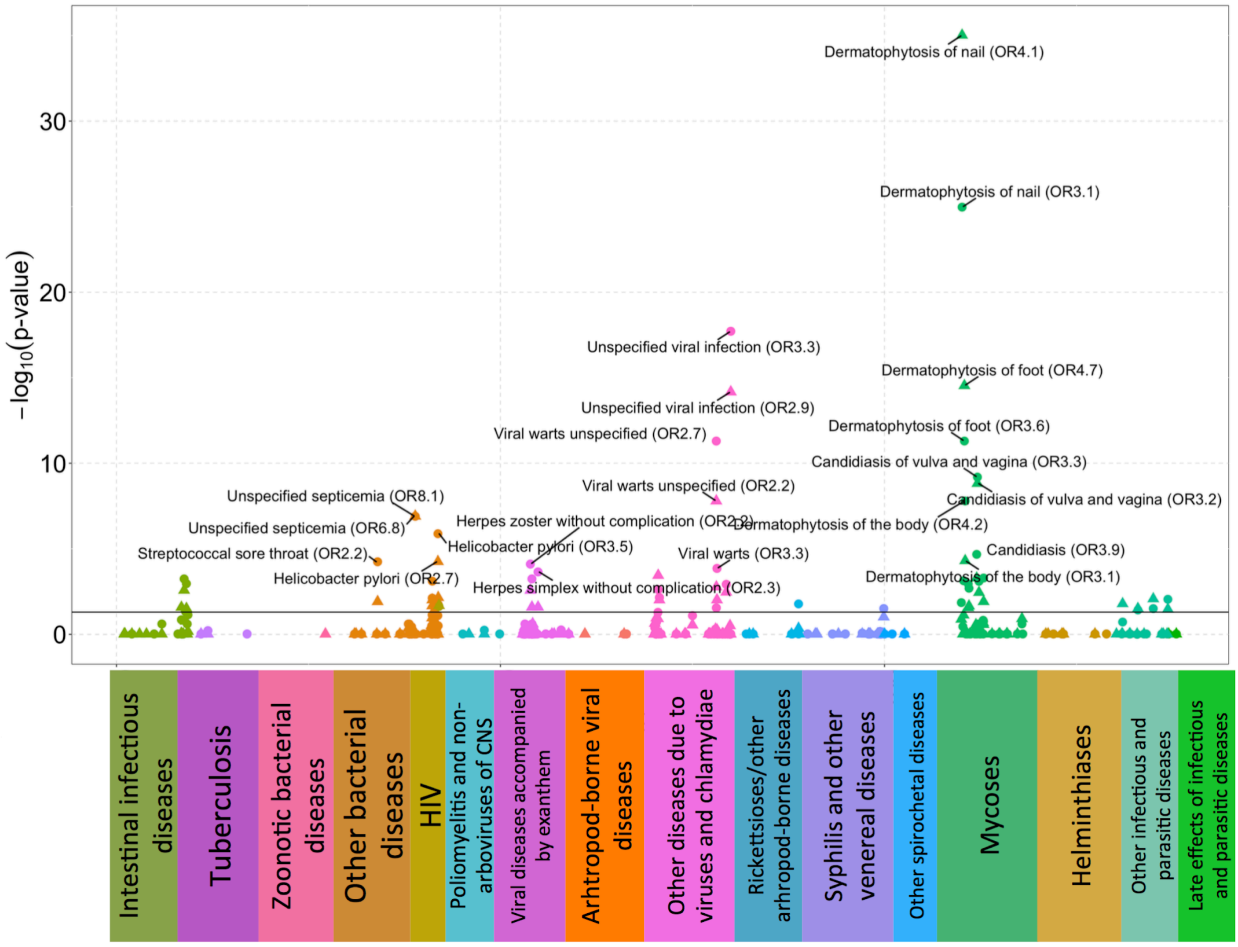


Figure 5.2. Manhattan plot style figure for the infection-wide association study of infectious disease diagnostic codes associated with type 2 diabetes in IWAS-b, or before the documented date of T2D diagnosis. The plot symbols represent the training (circles) and validation (triangles) sets. Each point represents an infectious disease ICD-9 diagnostic code and is labeled with the odds ratio presented in parentheses. The colors represent the different categories of infection represented in our study. The x-axis

indicates the ICD-9 diagnostic code arranged left to right in numeric order (0 to 139) and the y-axis indicates the -log10(p-value) of the adjusted logistic regression coefficient per infectious disease.

Infection complications of T2D

Figure 5.3 displays the distribution of p-values of association between each infection ICD-9 code and T2D for infections that occurred after the documented date of T2D diagnosis (IWAS-a). We identified 28 of 274 (11%) infections that met an FDR threshold of 5% in the training and validation sets in IWAS-a. The most significant (lowest FDR) results in IWAS-a were consistent with those of IWAS-b, except greater in magnitude of odds ratio: dermatophytosis of nail (OR 4·07, P = 5·4e-51) and foot (OR 4·65, P = 2·6E-37), unspecified septicemia (OR 11·15, P = 3·3e-18), candidiasis of mouth (OR 3·69, P = 8·2e-13) and skin and nails (OR 3·72, P = 4·5E-10) and vulva and vagina (OR 3·31, P = 4·0E-9), and *H. pylori* (OR 3·05, P = 4.7E-16) **(Supplementary Figure 5.2b)**.

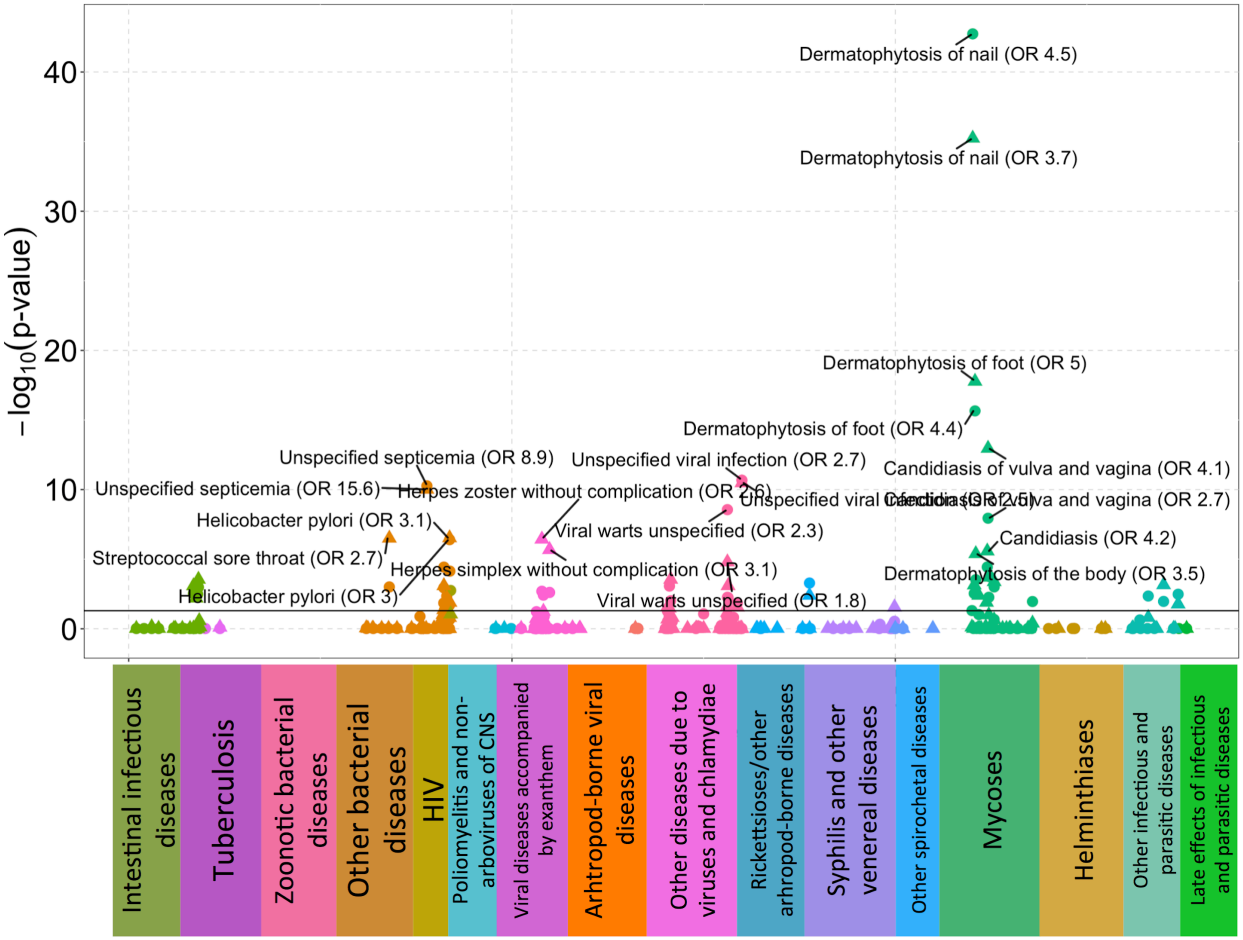


Figure 5.3. Manhattan plot style figure for the infection-wide association study of infectious disease diagnostic codes associated with type 2 diabetes in IWAS-a, or after the documented date of T2D diagnosis. The plot symbols represent the training (circles) and validation (triangles) sets. Each point represents an infectious disease ICD-9 diagnostic code and is labeled with the odds ratio presented in parentheses. The colors represent the different categories of infection represented in our study. The x-axis indicates the ICD-9 diagnostic code arranged left to right in numeric order (0 to 139), and the y-axis indicates the -log10(p-value) of the adjusted logistic regression coefficient per infectious disease.

The mean prevalence of the ICD-9 codes examined in the total population was low, at 0·09% **(Supplementary Figure 5.3, Supplementary Table 5.5)**. For the majority of identified ICD-9 codes, the odds ratio in IWAS-a was greater than that of IWAS-b, indicating a greater odds of developing the infection as a diabetic after than prior to the date of documented T2D **(Figure 5.4 [continued], Supplementary Table 5.6, Supplementary Figure 5.4)**. Of the 28 infections identified in IWAS-a, 23 (82%)

infections overlapped with the results for IWAS-b. The infections that were identified in IWAS-a and not identified in IWAS-b included intestinal infection due to *Clostridium difficile* (OR 4·98, P = 4·6e-5) and Lyme disease (OR 2·82, P = 4·9e-8). Conversely, the infections that were identified in IWAS-b and not included in IWAS-a included: methicillin susceptible staphylococcus aureus (OR 7·02, P = 7·0E-6), HIV (OR 5·24, P = 2·6e-5) and genital herpes (OR 3·5, P = 1·1e-6).

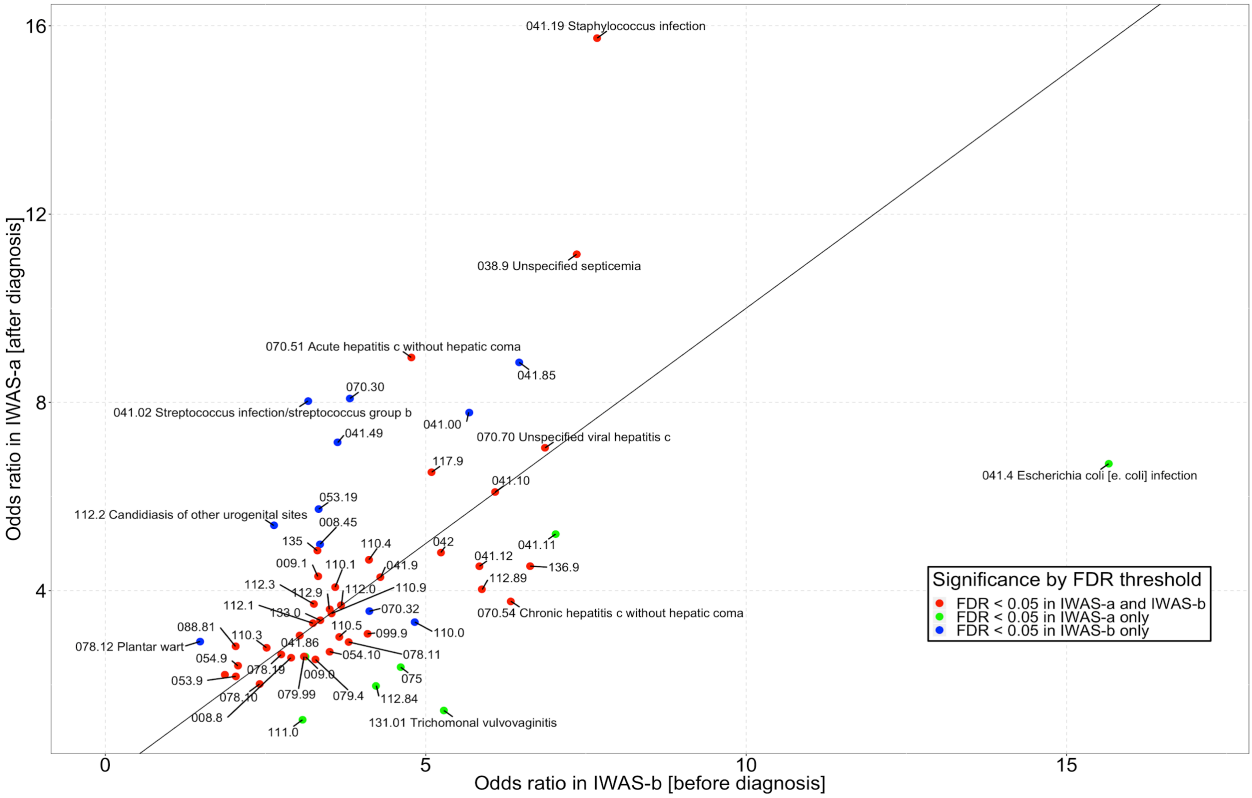


Figure 5.4. Plot of odds ratio (OR) in IWAS-a VS. OR in IWAS-b. Each point represents an ICD-9 diagnosis code and is labeled accordingly. Certain ICD-9 diagnostic codes are labeled with the corresponding name of the infectious disease; the full list of infectious disease names corresponding to the labeled ICD-9 codes are presented in **Supplementary Table 5.6**.

In the granular 1-year time interval results, there was a greater odds of association with T2D within the period of one-to-two years prior the date of documented T2D for viral warts, herpes simplex, and Helicobacter pylori **(Supplementary Figure 5.5)**. For methicillin resistant *Staphylococcus aureus* and candidiasis, there is a greater odds of association to T2D in proportion to decreased time from infection to the date of documented diagnosis, which increases in odds ratio magnitude for 1 year after the date of T2D diagnosis, and then declines for the two years that follow **(Supplementary Figure 5.5)**.

Mendelian randomization assessment of infectious agents and T2D

Table 5.2 shows the results of MR analyses. As shown, all these analyses yielded results that were not even nominally statistically significant, with one exception. For Hepatitis B, we identified a suggestive estimate of 0·1413 (SE = 0·053, P = 0·008) for association with T2D.

Table 5.2. Summary of Mendelian randomization derived estimates.

Exposure	Outcome	IV, n	SNP(s)	Estimate (β)	Standard error (SE)	P-value
Antibody level in response to infection, unit increase	T2D	1	rs4812712	0.055	0.112	0.623
Antibody level in response to infection, unit increase	HbA1c	1	rs4812712	-0.011	0.052	0.838
Chronic Hepatitis C	T2D	1	rs8099917	-0.0353	0.024	0.136
Chronic Hepatitis C	HbA1C	1	rs8099917	-0.00012	0.0006	0.985
Chronic Hepatitis C	Fasting insulin	1	rs8099917	-0.0029	0.0075	0.703
Helicobacter pylori	T2D	2	rs368433, rs10004195	0.032	0.089	0.722
Helicobacter pylori	HbA1c	2	rs368433, rs10004195	-0.0094	0.0095	0.318
Helicobacter pylori	Fasting insulin	2	rs368433, rs10004195	-0.019	0.023	0.399
Hepatitis B	T2D	1	rs9268652	0.1413	0.053	0.008
Hepatitis B	HbA1C	1	rs9268652	0.0295	0.014	0.040
Hepatitis B	Fasting insulin	1	rs9268652	0.0202	0.016	0.203
HIV-1 susceptibility	T2D	1	rs4878712	-0.036	0.045	0.429
HIV-1 susceptibility	HbA1C	1	rs4878712	0.0076	0.013	0.560
HIV-1 susceptibility	Fasting insulin	1	rs4878712	-0.019	0.014	0.185
Yeast infection	T2D	2	rs2251260, rs716578	-0.156	0.230	0.499
Yeast infection	HbA1C	2	rs2251260, rs716578	0.070	0.062	0.259
Yeast infection	Fasting insulin	2	rs2251260, rs716578	-0.056	0.153	0.712

DISCUSSION

Our study extends the current state of knowledge on the associations between infection and T2D, through the lens of infection as a candidate risk factor and as a complication for T2D. First, we document 252 and 274 infectious diseases that occur in a large cohort of 172,000 patients with T2D and matched controls before and after documented diagnosis of T2D. Second, we comprehensively and systematically assess each infectious disease in association with T2D. In doing so, we have corroborated previously reported associations such as a predisposition to fungal infections including Candida and Herpes simplex virus[120,140] and T2D incidence in HIV-infected patients[141], and identified tentatively novel associations (to our knowledge), including methicillin-susceptible *Staphylococcus aureus* and acariasis associated with T2D risk and Lyme disease associated with T2D as a complication. Previously reported findings on the relationship between *Helicobacter pylori* and T2D are conflicting[142,143]. In our investigation, IWAS-b identified an OR of 3·1, in concordance with a previously reported finding that individuals seropositive for the infection were 2·7 times more likely to develop diabetes than seronegative individuals[144].

Third, we extend our observational analysis with time-varying granularity by examining infection-T2D associations in 12-month intervals covering a time period of three years before and after documented diagnosis of T2D. We found that 23 (or 74% of the IWAS-b) infections were also found in associations after date of diagnosis (IWAS-a). Furthermore, we identified that for most of the infections found in both IWAS-a and IWAS-b, the magnitude of odds ratio was greater in IWAS-a, indicating an increased odds of infection after the T2D diagnosis was formally recorded. As in other

observational investigations, our study is susceptible to reverse causality. By examining infection both before and after disease onset, we describe the extent of the lack of clarity between the direction of association between infection and T2D. Fourth, we assess findings from our observational investigation in a MR framework in order to test for risk of infection on T2D and glycemic traits.

Since many of the infections had increased risks both before and after T2D diagnosis, this could implicate that some infections may be both risk factors for and complications of T2D. Alternatively, it may be indicative that a diagnosis code for T2D appears up to many years after the disease onset. Future studies that use convenience samples from administrative data should recognize the challenges of identifying incident cases in their analyses ("incident" cases may be prevalent cases). With MR, we investigated the potential directional effect between infection and T2D, assuming that some individuals harbor genetic variants that increase their susceptibility to infection. MR provides limited evidence for a causal association between infection and documented T2D, supporting the hypothesis that infection may play a far greater role as a complication of T2D rather than a risk factor.

A limitation of our study is potential misclassification of documented date of diagnosis. However, we tried to mitigate this limitation by requiring all cases to have a T2D specific ICD-9 code, without anti-T2D medication, and without a laboratory test result for fasting glucose, non-fasting glucose, or HbA1C that may be indicative of T2D prior to the first documented T2D ICD-9 code. Another limitation includes the low prevalence of recorded infection, which could affect our statistical power to detect associations. It is possible that many infections are not reported in these insurance data. We cannot also

exclude the possibility that infections are more likely to be diagnosed and reported in these claims data for patients who have already an established diagnosis of T2D, while clinicians may not pursue as rigorously some infectious disease diagnoses if patients are not known to be diabetic. Lastly, for the infections we identified in IWAS-b, most (n=26) did not have available publicly available GWAS summary statistics to assess the risk of infection using MR, and for the infections that did, there were a low number of genetic instruments (or variants) that could be used as a proxy for the infection. Therefore, even though the current MR analyses are largely "negative", the entire space of infections that emerged in IWAS-b (n=31) is yet to be explored and should not be ruled out.

In conclusion, our study documents the magnitude of associations between the full spectrum of infectious diseases and T2D over a six-year time period and uses MR to detect potential causal effects. Future investigation is warranted to study why individuals with T2D are more susceptible to particular infections, both as a risk factor and as a complication of the disease, and to identify patient populations who may have greater susceptibility to infection in order to improve preventative care.

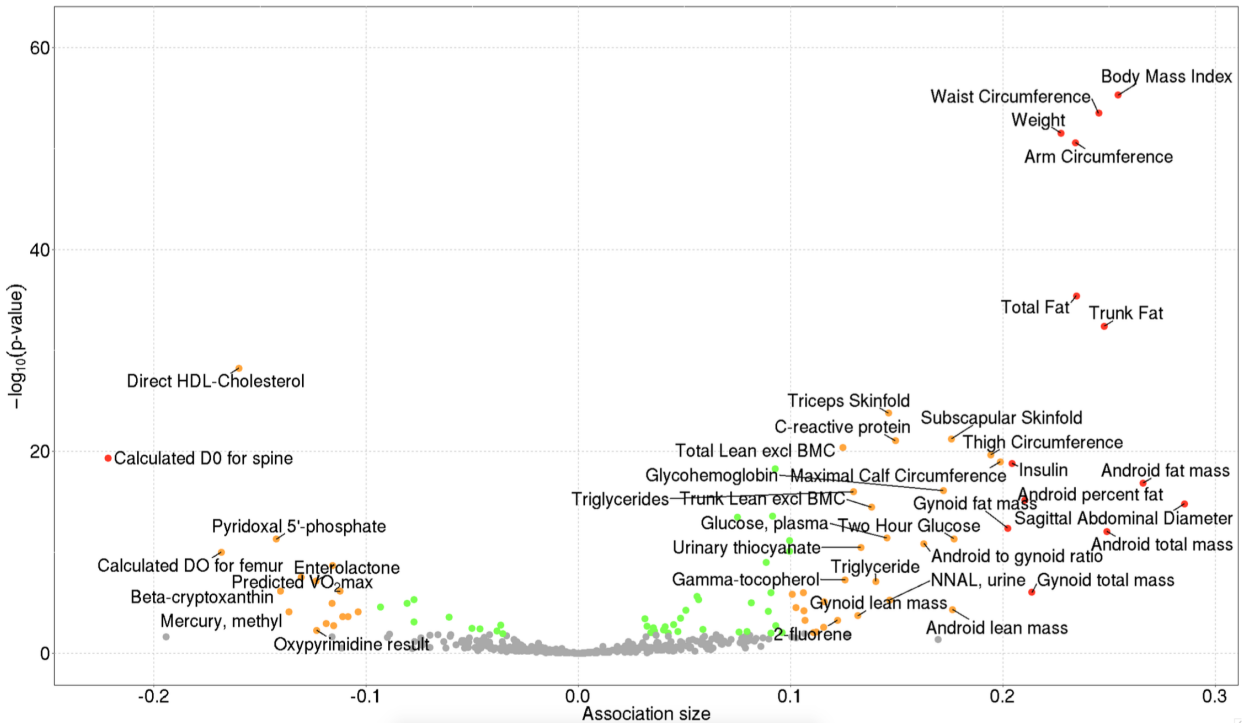
6
Conclusion

In conclusion, we demonstrate high-throughput informatics frameworks for studying familial influences in disease risk without costly direct case genotyping and leverage genetic variation to examine putative causal relationships from our observational studies. First, our family history-based approaches advance our understanding of how family history affects inherited traits and environmental factors, demonstrates the combined impact of multiple family histories on disease risk and risk factors, and disentangles potential environmental contributions from genetic influences of familial risk. Second, we demonstrate the integration of Mendelian randomization to assess causality in associations from findings in an observational investigation, presenting a data-driven map of risk factors and/or complications of type 2 diabetes. Our findings underscore the public health impact and utility of family history and Mendelian randomization as tools for

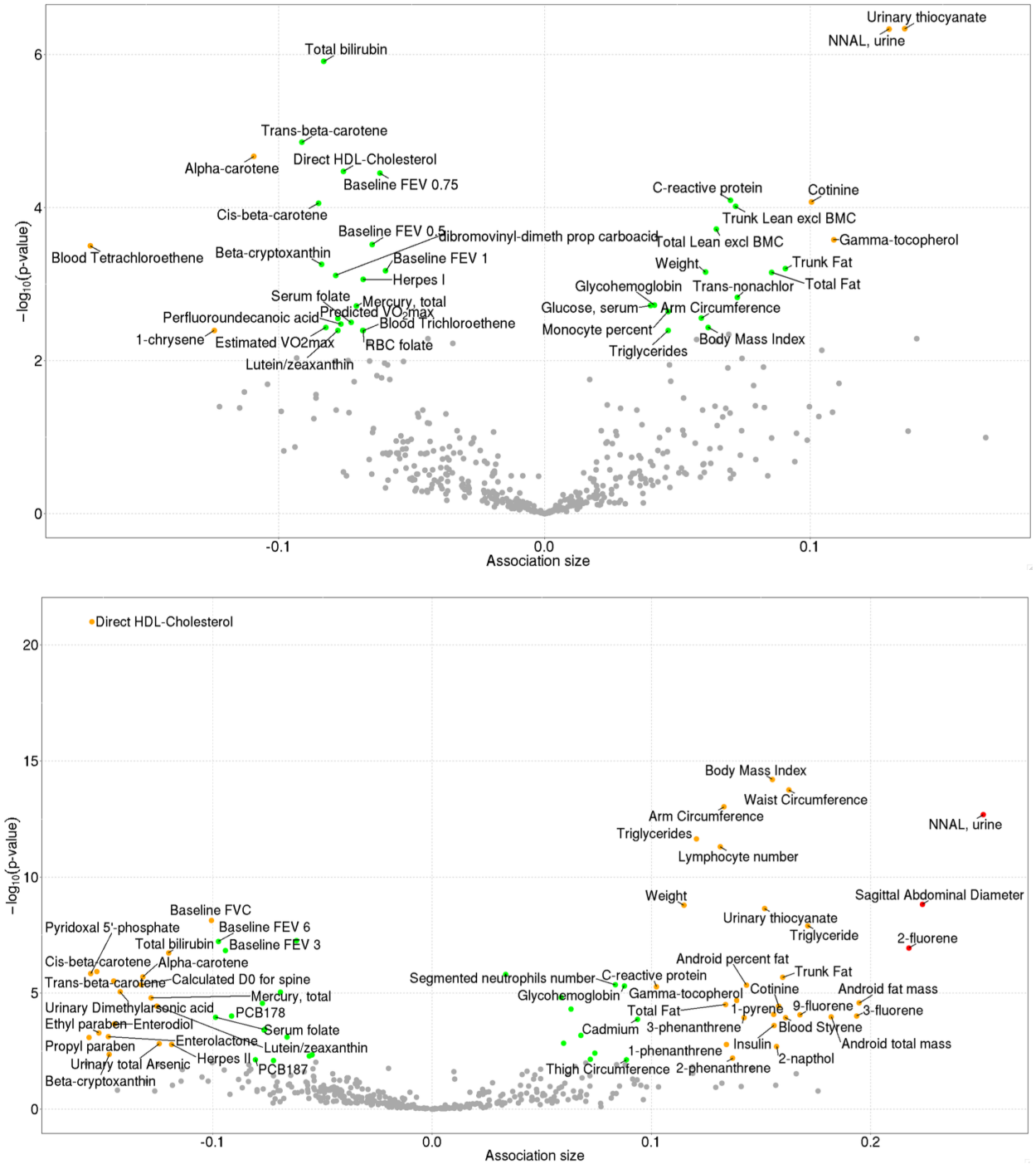
distinguishing potential areas for early detection, prevention, or therapeutic intervention of chronic disease.

We illustrate in our studies methods for driving discovery and for extracting insight on mechanisms underlying associations with disease by analyzing troves of clinical and genomics data contained within large-scale resources such as the UK Biobank, NHANES, and health insurance claims datasets. The scale and breadth of these resources, coupled with machine learning approaches, can yield new insights for enhancing our understanding of biomedicine, informing on public health approaches, and assisting in therapeutic discovery. In light of the potential, however, data science and machine learning techniques are prone to a multitude of challenges and limitations, such as those that have been mentioned in this thesis. Future research that I aspire to work on include development of machine learning methods for causal learning incorporated with genetics-based Mendelian randomization analyses for characterizing large-scale health and omics datasets. Findings from such methodology can, potentially, pave the way for new insights on disease aetiology.

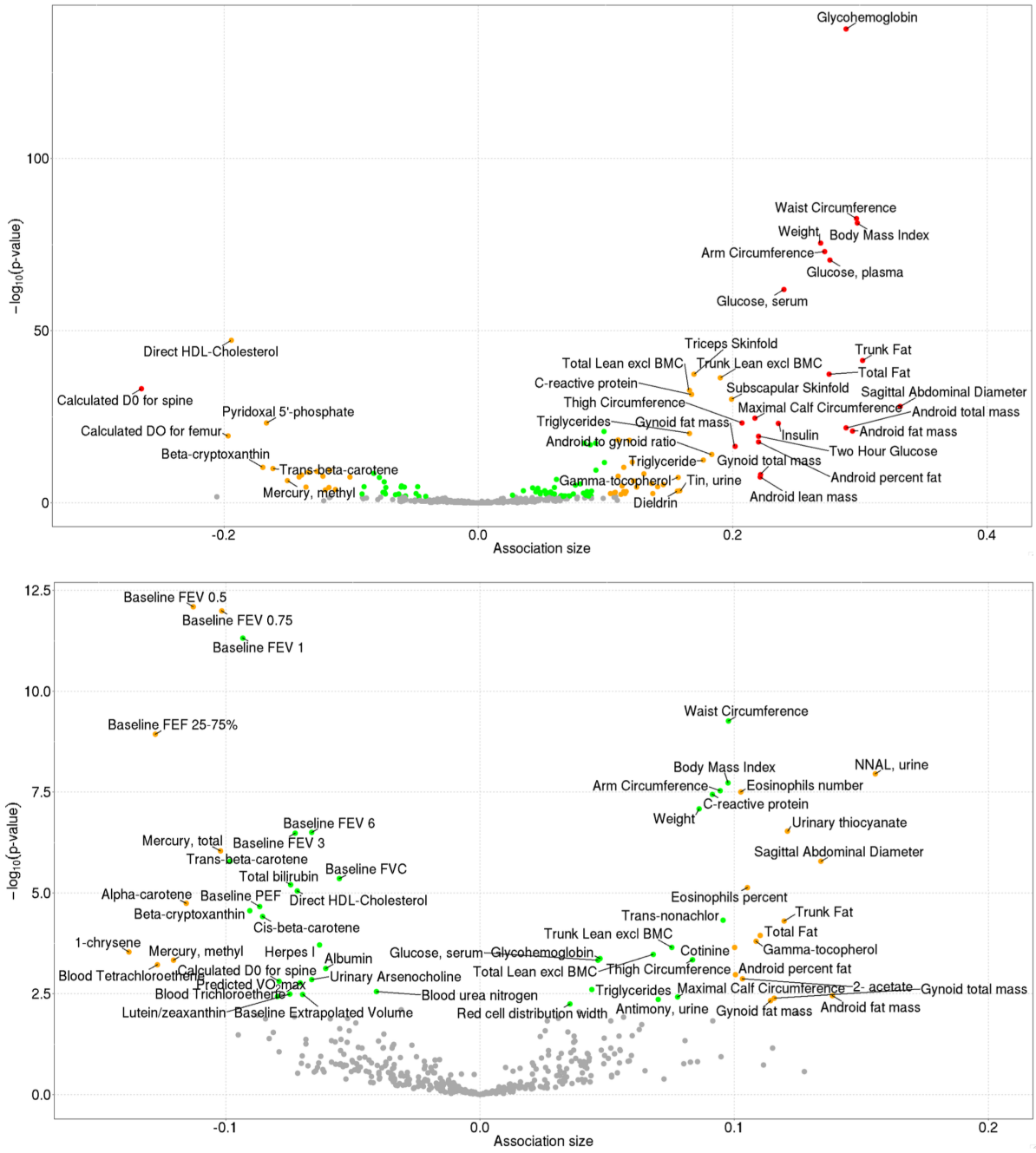
Appendix



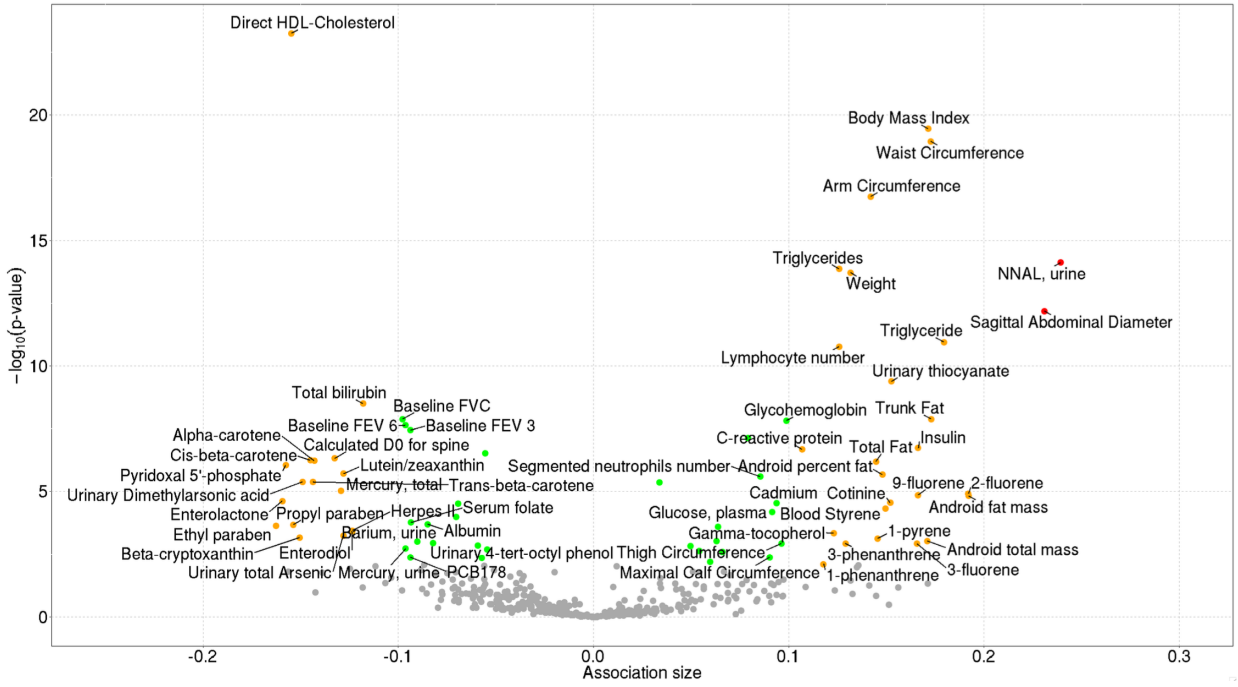
Supplementary Figure 1.1a. Volcano plot results for meta-analysis in individuals without the respective disease. 457 CEQT association sizes versus -log10(p-value) for family history of diabetes (**Supplementary Figure 1.1a**), asthma (**Supplementary Figure 1.1b**), and coronary heart disease (**Supplementary Figure 1.1C**) in individuals without the respective disease, adjusting for age, sex, and race, in 1999-2014 National Health and Nutrition Examination Survey (NHANES). Green, orange, and red points represent traits that met an FDR threshold of 5%. Orange and red points represent traits with an absolute value of association size greater than or equal to 0.10 and 0.20, respectively. All labeled points are traits that met an FDR of 5% and have an absolute value of association size greater than or equal to 0.12 in A, 0.02 in B, and 0.08 in C.



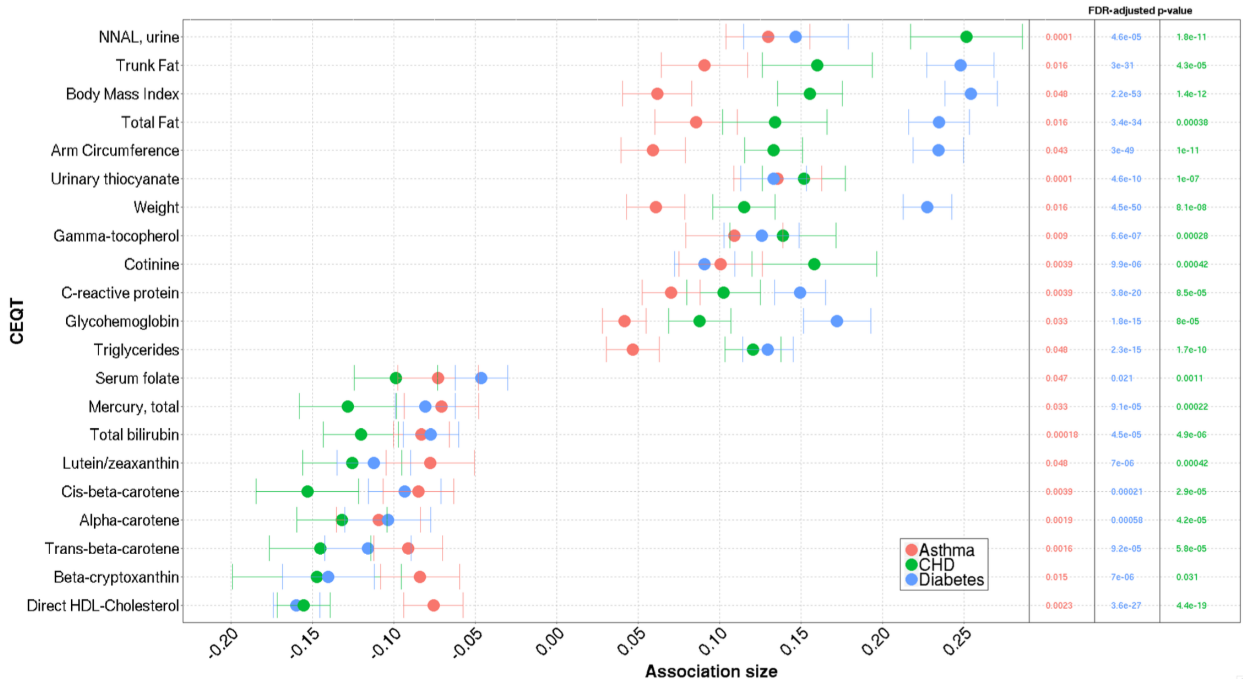
Supplementary Figure 1.1b for asthma (top) and Supplementary Figure 1.1c for coronary heart disease (bottom). 457 CEQT association sizes versus -log10(p-value) for family history of diabetes (**Supplementary Figure 1.1a**), asthma (**Supplementary Figure 1.1b**), and coronary heart disease (**Supplementary Figure 1.1C**) in individuals without the respective disease, adjusting for age, sex, and race, in 1999-2014 National Health and Nutrition Examination Survey (NHANES).



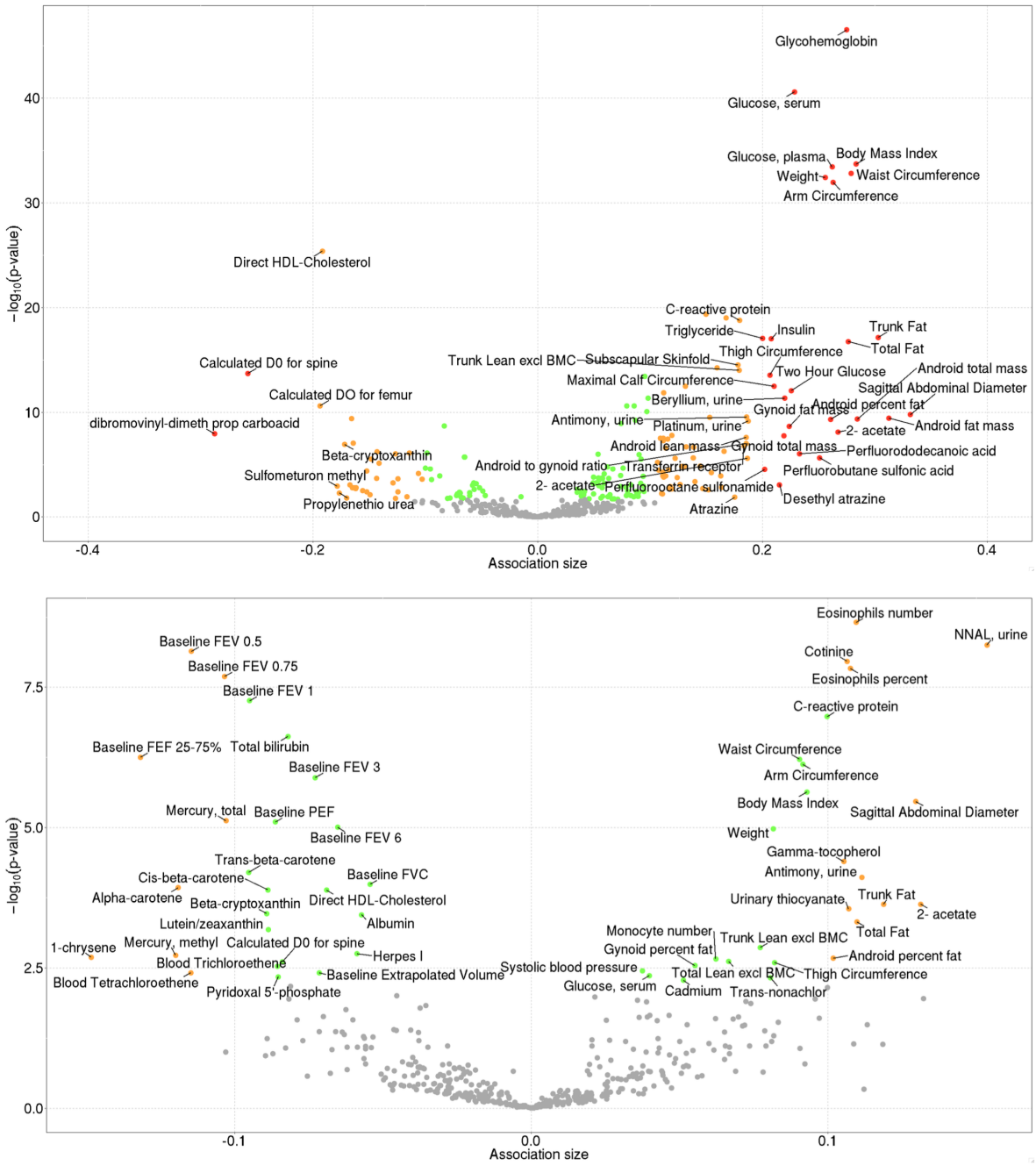
Supplementary Figures 1.2a and b. Volcano plot results for meta-analysis in all individuals. 457 CEQT association sizes versus -log10(p-value) for family history of diabetes (top) and asthma (bottom) in all individuals (with and without respective disease), adjusting for age, sex, and race, in 1999-2014 National Health and Nutrition Examination Survey (NHANES). Green, orange, and red points represent traits that met an FDR threshold of 5%. Orange and red points represent traits with an absolute value of association size greater than or equal to 0.10 and 0.20, respectively. All labeled points are traits that met an FDR of 5% and have an absolute value of association size greater than or equal to 0.15 in A and 0.02 in B.



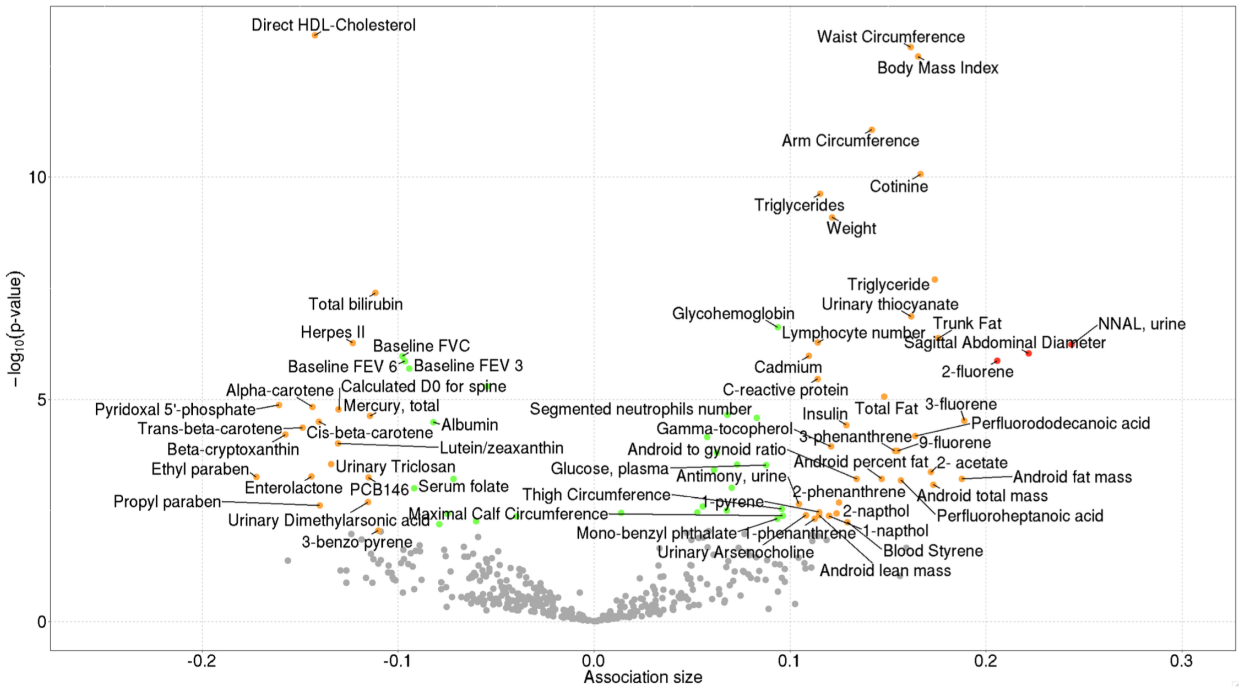
Supplementary Figures 1.2c. Volcano plot results for meta-analysis in all individuals. 457 CEQT association sizes versus -log10(p-value) for coronary heart disease in all individuals (with and without respective disease), adjusting for age, sex, and race, in 1999-2014 National Health and Nutrition Examination Survey (NHANES). Green, orange, and red points represent traits that met an FDR threshold of 5%. Orange and red points represent traits with an absolute value of association size greater than or equal to 0.10 and 0.20, respectively. All labeled points are traits that met an FDR of 5% and have an absolute value of association size greater than or equal to 0.08.



Supplementary Figure 1.3. Shared CEQTs from meta-analysis. From the meta-analysis, shared CEQTs associated with family histories of diabetes (shown in blue), asthma (red), and coronary heart disease (green) in individuals without the respective disease in 1999-2014 National Health and Nutrition Examination Survey (NHANES). All CEQTs displayed achieved an FDR threshold of 5%. All models are adjusted for age, sex, and race. The FDR-adjusted p-value for each point is displayed to the right (in red for asthma, blue for diabetes, and green for coronary heart disease).

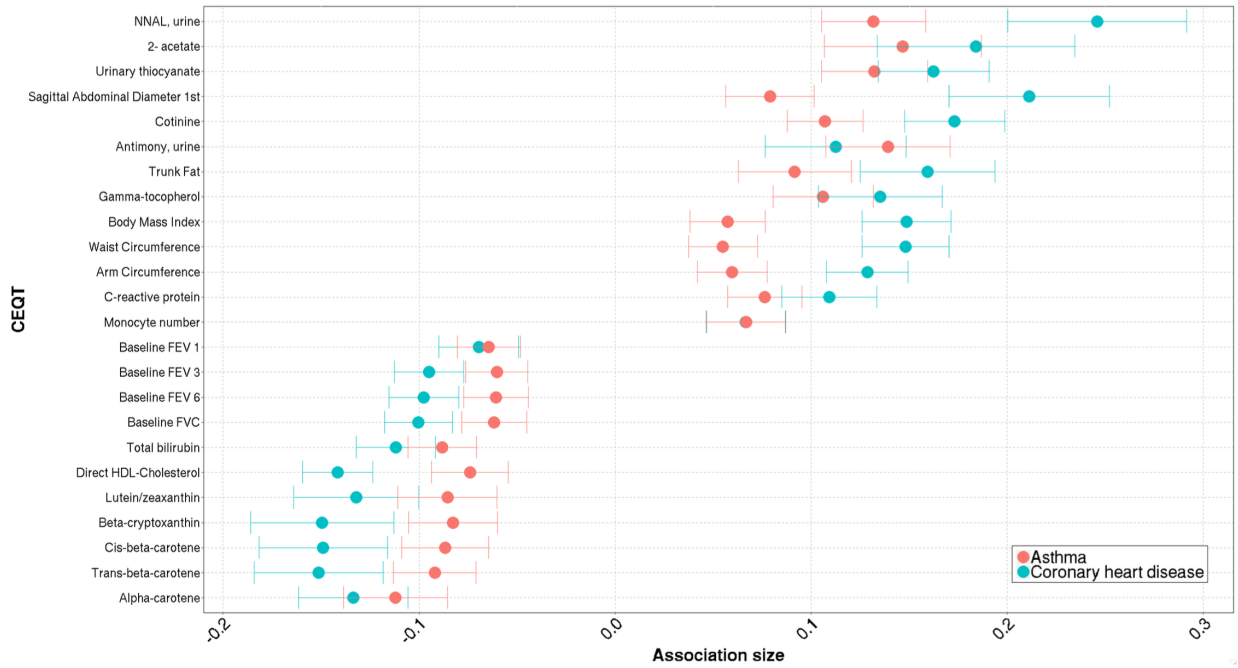


Supplementary Figure 1.4a and b. Volcano plot results for pooled analysis in all individuals. 457 CEQT association sizes versus -log10(p-value) for family history of diabetes (**top**) and asthma (**bottom**) in all individuals (with and without respective disease), adjusting for age, sex, and race, in 1999-2014 National Health and Nutrition Examination Survey (NHANES). Green, orange, and red points represent traits that met an FDR threshold of 5%. Orange and red points represent traits with an absolute value of association size greater than or equal to 0.10 and 0.20, respectively. All labeled points are traits that met an FDR of 5% and have an absolute value of association size greater than or equal to 0.17 in A and 0.02 in B.

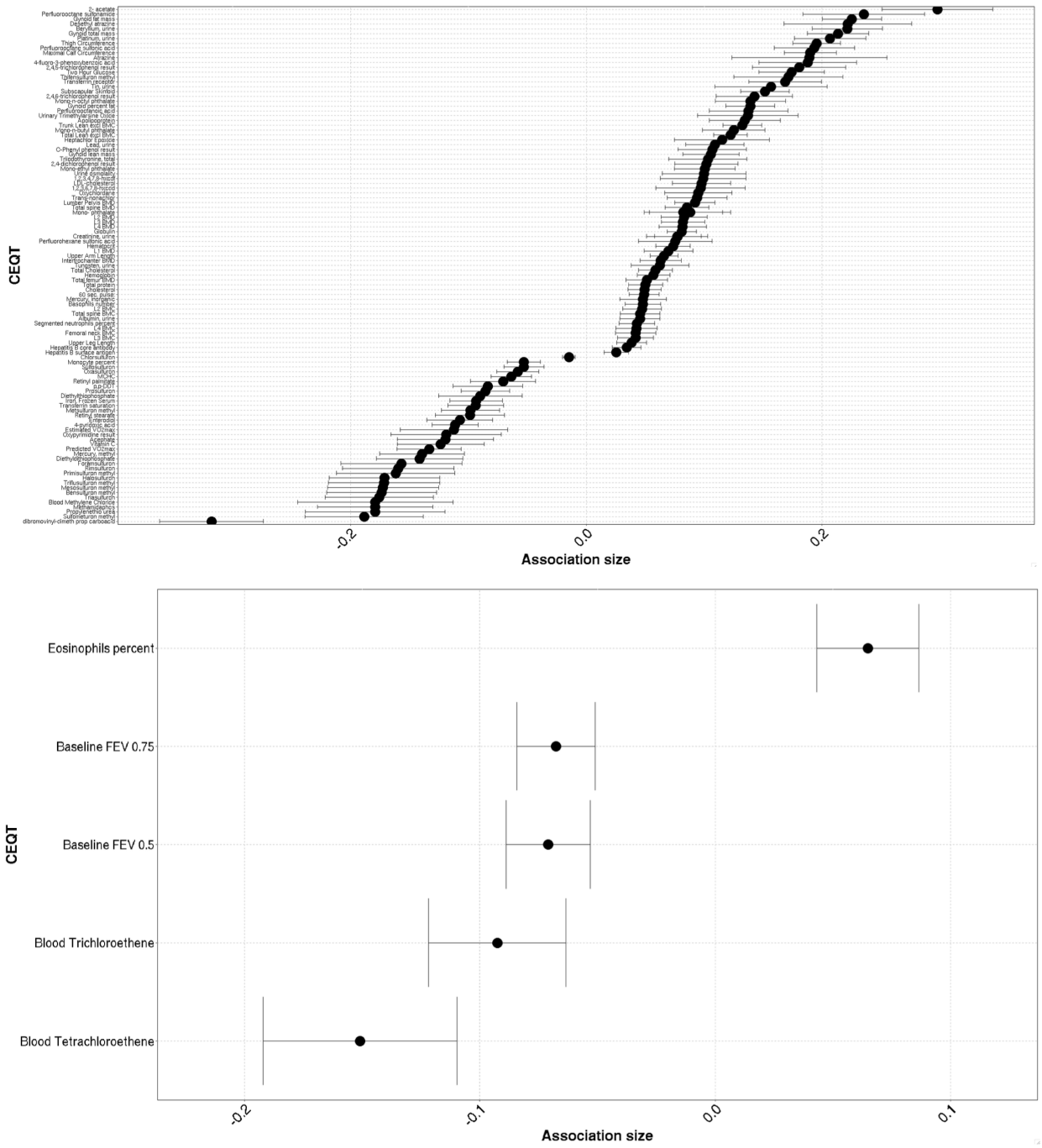


Supplementary Figure 1.4c. Volcano plot results for pooled analysis in all individuals. 457 CEQT association sizes versus -log10(p-value) for family history of coronary heart disease (**continued**) in all individuals (with and without respective disease), adjusting for age, sex, and race, in 1999-2014 National Health and Nutrition Examination Survey (NHANES). All labeled points are traits that met an FDR of 5% and have an absolute value of association size greater than or equal to 0.08 in Supplementary Figure 1.4C.

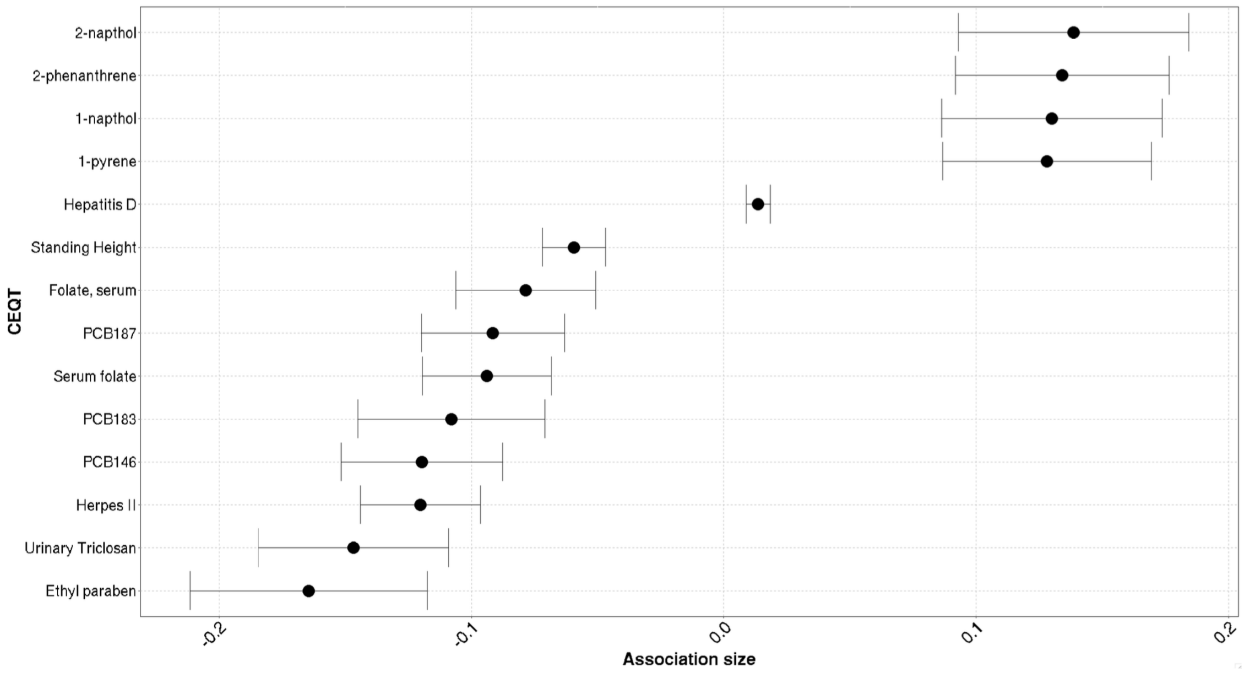
Supplementary Figure 1.5a and b. Shared CEQTs associated with family histories of diabetes, asthma, and coronary heart disease is shown in individuals without respective disease, adjusted by age, sex, and race. All CEQTs displayed achieved FDR of 5%. We identified 60 CEQTs shared between family histories of coronary heart disease and diabetes (**top, Supplementary Figure 1.5A**), 23 between family histories of asthma and diabetes (**bottom, Supplementary Figure 1.5B**), and 24 between family histories of asthma and coronary heart disease (**Supplementary Figure 1.5C, Continued**).



Supplementary Figure 1.5c. Shared CEQTs associated with family histories of asthma and coronary heart disease is shown in individuals without respective disease, adjusted by age, sex, and race. We found 24 between family histories of asthma and coronary heart disease (**Supplementary Figure 1.5C, Continued**).



Supplementary Figure 1.6a and b. Unique CEQTs associated (FDR of 5%) with one family history, and not with the other two family histories is shown in individuals without respective disease, adjusted by age, sex, and race. We identified 109 CEQTs associated with a family history of diabetes and not with family histories of asthma or CHD (**top, Supplementary Figure 1.6A**), 5 CEQTs associated with a family history of asthma and not with family histories of diabetes or CHD **(bottom, Supplementary Figure 1.6B**), and 14 CEQTs associated with a family history of CHD and not with diabetes or asthma (**Supplementary Figure 1.6C, Continued**).



Supplementary Figure 1.6c. Unique CEQTs associated (FDR of 5%) with one family history, and not with the other two family histories is shown in individuals without respective disease, adjusted by age, sex, and race. We identified 14 CEQTs associated with a family history of CHD and not with diabetes or asthma.

Supplementary Table 2.1. Estimates of adjusted odds ratios (95% CI) for CVD and diabetes according to family history, NHANES 2007-2018. All models are adjusted by the factors listed below (family history, age, gender, race/ethnicity, BMI, income-to-poverty-ratio, and education).

	CVD diagnosis			**Diabetes diagnosis**		
Reported family history of:	**CVD and no diabetes**	**Diabetes and no CVD**	**CVD and diabetes**	**CVD and no diabetes**	**Diabetes and no CVD**	**CVD and diabetes**
Family history (Yes vs No†)	3.04 (2.65,3.5)	1.39 (1.23,1.56)	3.37 (2.74,4.14)	2.49 (2.14,2.9)	3.03 (2.69,3.42)	3.65 (3.06,4.36)
Age (20-59†) 60+	6.83 (5.96,7.83)	8.38 (7.19,9.77)	7.32 (6.31,8.48)	3.8 (3.31,4.36)	4.28 (3.80,4.81)	4.04 (3.49,4.69)
Sex (Females†) Males	1.79 (1.55,2.08)	1.83 (1.6,2.1)	1.79 (1.54,2.09)	1.49 (1.32,1.69)	1.62 (1.44,1.82)	1.52 (1.33,1.73)
Race/Ethnicity (Non-Hispanic White†) Hispanic	0.57 (0.48,0.67)	0.54 (0.46,0.65)	0.53 (0.44,0.64)	1.32 (1.14,1.53)	1.31 (1.16,1.48)	1.29 (1.12,1.5)
Non-Hispanic Black	0.85 (0.73,0.99)	0.94 (0.81,1.08)*	0.82 (0.69,0.97)	1.54 (1.32,1.8)	1.57 (1.4,1.76)	1.46 (1.23,1.73)
Other	0.87 (0.66,1.14)*	0.91 (0.71,1.17)*	0.83 (0.61,1.11)*	1.84 (1.48,2.28)	1.91 (1.59,2.29)	1.86 (1.49,2.31)
BMI (<25†) 25-29.9	1.09 (0.94,1.27)*	1.03 (0.88,1.20)*	1.15 (0.97,1.38)*	1.79 (1.48,2.17)	1.75 (1.5,2.04)	1.8 (1.45,2.24)
BMI 30+	1.48 (1.26,1.74)	1.44 (1.25,1.67)	1.43 (1.21,1.69)	5.59 (4.48,6.98)	4.36 (3.73,5.08)	5.3 (4.18,6.72)
Income to poverty ratio (>3.5†) 1.3-3.49	1.75 (1.49,2.05)	1.77 (1.45,2.16)	1.76 (1.46,2.12)	1.25 (1.03,1.52)	1.17 (1.02,1.33)	1.27 (1.03,1.57)
<1.3	2.03 (1.71,2.41)	1.96 (1.68,2.29)	2.06 (1.69,2.52)	1.43 (1.21,1.7)	1.20 (1.06,1.37)	1.40 (1.17,1.67)
Education (High school completion or greater†) Less than high school	1.32 (1.12,1.55)	1.43 (1.23,1.65)	1.34 (1.12,1.61)	1.50 (1.32,1.71)	1.47 (1.32,1.63)	1.48 (1.3,1.69)

*† : Reference group |*Values marked with an asterisk do not meet an FDR significance threshold of 5%.*

Supplementary Table 3.1. Fields used to ascertain current disease status.

Family history phenotypes	Disease phenotypes
Prostate cancer	Self-reported cancer Data-field 20001
Severe depression	Self-reported non-cancer Data-field 20002. Depression (#1286)
Parkinson's Disease	Self-reported non-cancer Data-field 20002. Parkinson's. #1262
Alzheimer's disease/dementia	Self-reported non-cancer Data-field 20002. Dementia/Alzheimers/Cognitive impairment #1263
Diabetes	"Diabetes diagnosed by doctor" Medical Conditions. Touchscreen. (Yes or No)
High blood pressure	Vascular/heart problems diagnosed by doctor (Data-Field #6150). High blood pressure.
Chronic bronchitis/emphysema	Self-reported non-cancer Data-field 20002. Emphysema/Chronic bronchitis. #1113
Breast cancer	Self-reported cancer Data-field 20001
Bowel cancer	Self-reported cancer Data-field 20001
Lung cancer	Self-reported cancer Data-field 20001
Stroke	Self-reported non-cancer Data-field 20002. Ischaemic #1583 and stroke #1081.
Heart Disease	Vascular/heart problems diagnosed by doctor (Data-Field #6150).

Supplementary Table 3.2. Demographic breakdown of non-adopted and adopted cohorts in the UK Biobank.

	Non-adopted cohort	Adopted cohort
N	494,513	6,347
Age (Mean [95% CI])	56.5 [48.4, 64.6]	56.2 [47.6, 64.8]
% Female	54.4	52.9
Race* (%)		
White	94.3	91.7
Mixed	0.60	2.33
Asian or Asian British	1.95	1.30
Black or Black British	1.60	2.27
Chinese	0.31	0.57
Other ethnic group	0.91	1.09
Do not know	0.04	0.35
Prefer not to answer	0.32	0.38

*White includes British, Irish, and Any other white background. Mixed includes White and Black Caribbean, White and Black African, White and Asian, and Any other mixed background. Asian or Asian British include Indian, Pakistani, Bangladeshi, any Any other Asian background. Black or Black British includes Caribbean, African, and Any other Black background.

Supplementary Table 3.3. Disease and family history of disease breakdown of UK Biobank adopted and biological cohorts.

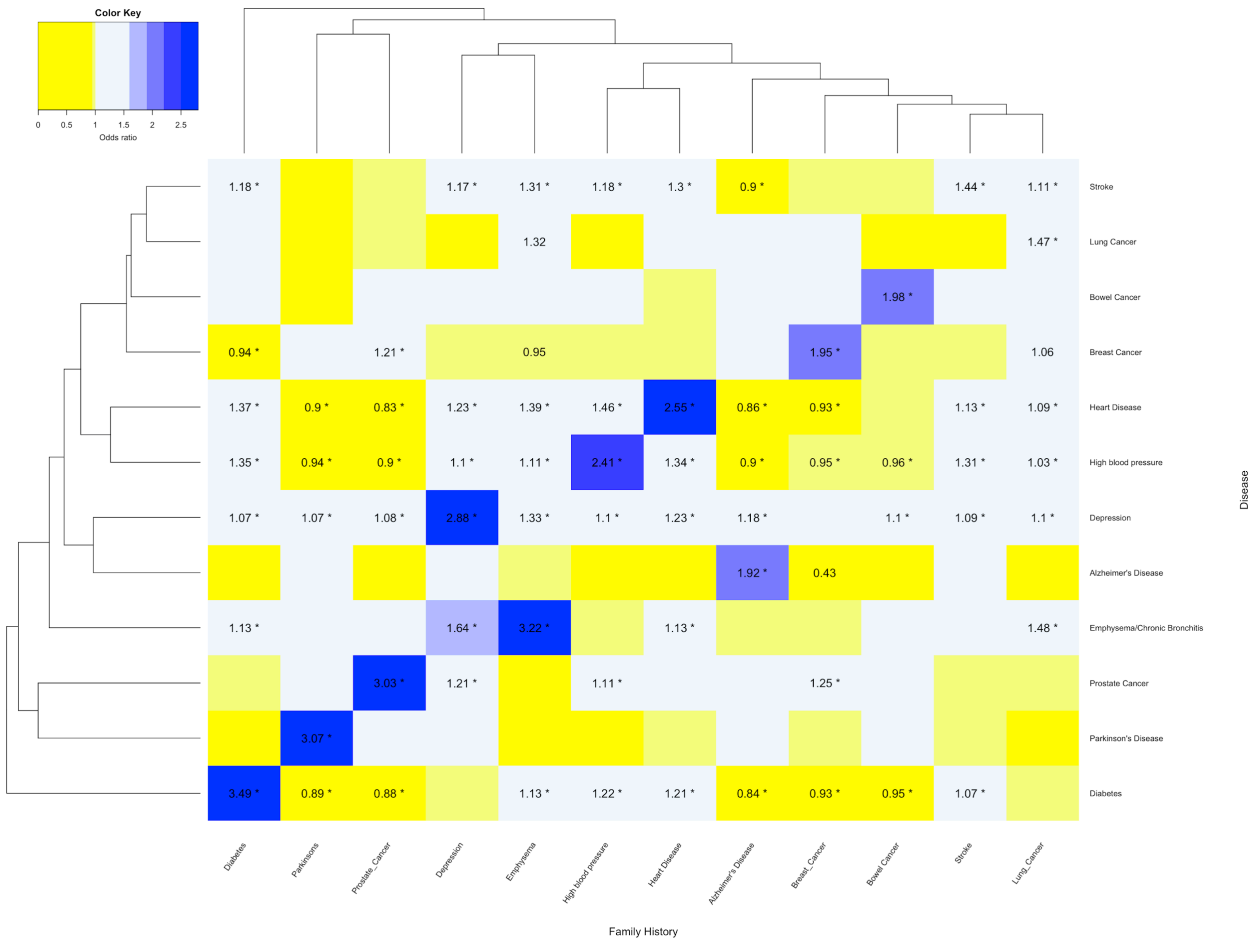
	Disease		Family history	
	Biological cohort	Adopted cohort	Biological cohort	Adopted cohort
Diabetes	27,024	505	109,665	876
Prostate Cancer	3,754	35	39,655	341
Breast Cancer	11,578	134	53,210	519
Bowel Cancer	977	<20	55,459	540
Depression	29,872	480	66,197	652
Parkinson's Disease	890	<20	20,199	211
Alzheimer's Disease	137	<20	61,681	633
Emphysema	7018	146	79,345	984
Stroke	8061	134	133,426	1294
High blood pressure	138,122	1874	241,757	2013
Heart disease	16,492	279	217,599	2164
Lung Cancer	315	<20	62,980	691

*Sample sizes less than 20 are indicated by "<20".

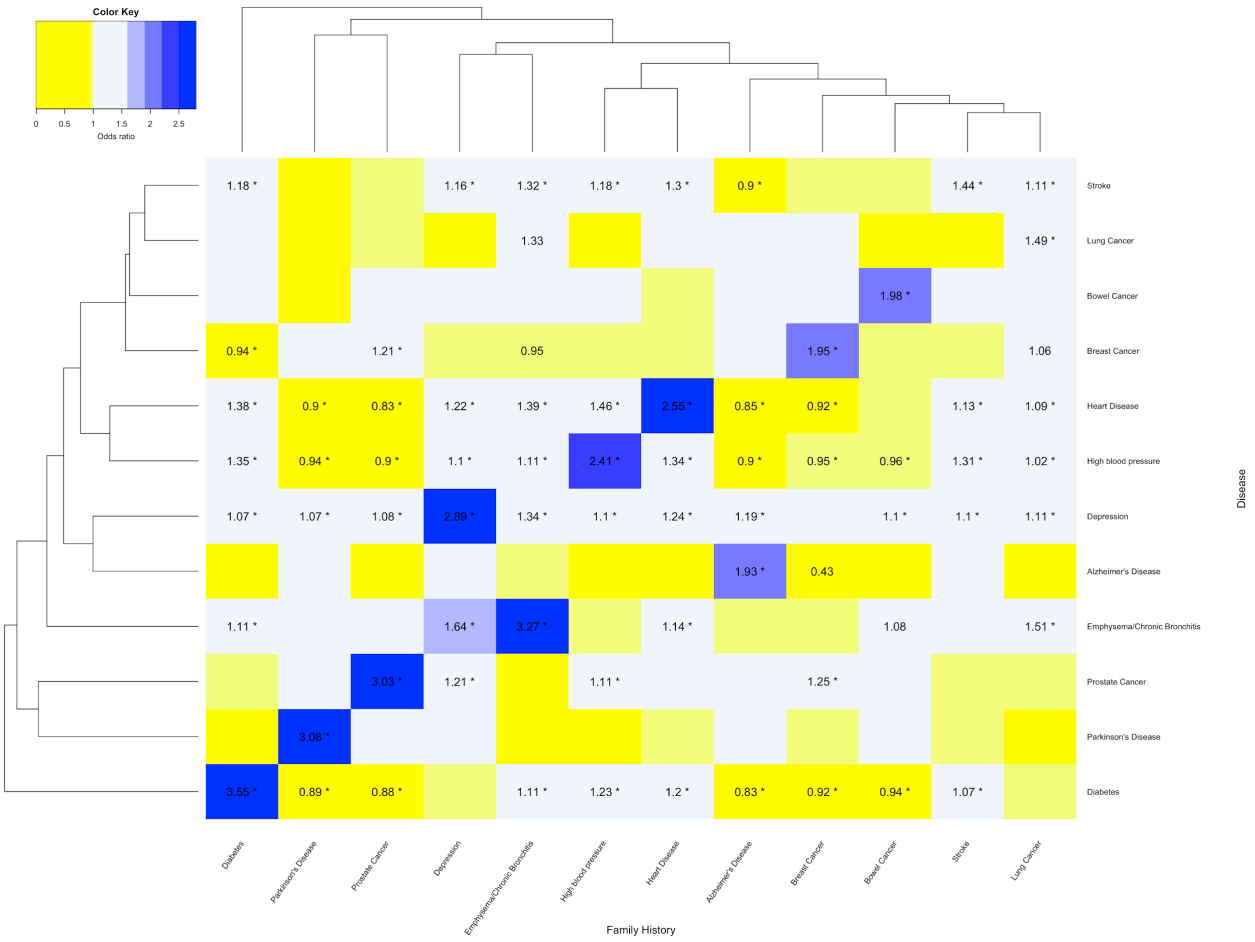
Supplementary Table 3.4. Parental and sibling family history of disease breakdown of UK Biobank adopted and biological cohorts.

	Biological			Adopted		
Family history	Paternal	Maternal	Sibling	Paternal	Maternal	Sibling
Diabetes	38,969	41,440	23,284	271	477	20
Prostate Cancer	33,761	-	5,903	309	-	27
Breast Cancer	-	37,236	15,987	-	120	95
Bowel Cancer	24,005	22,158	7,909	237	<20	33
Depression	14,667	28,925	20,624	127	454	150
Parkinson's Disease	10,624	7,322	2,055	118	<20	<20
Alzheimer's Disease	19,326	38,322	1,757	182	<20	20
Emphysema	45,524	23,379	6,565	580	141	71
Stroke	58,536	56,184	8,427	580	129	60
High blood pressure	60,950	56,184	34,533	559	1769	221
Heart disease	109,065	58,594	15,665	1008	273	130
Lung Cancer	37,272	16,184	7,013	398	<20	70

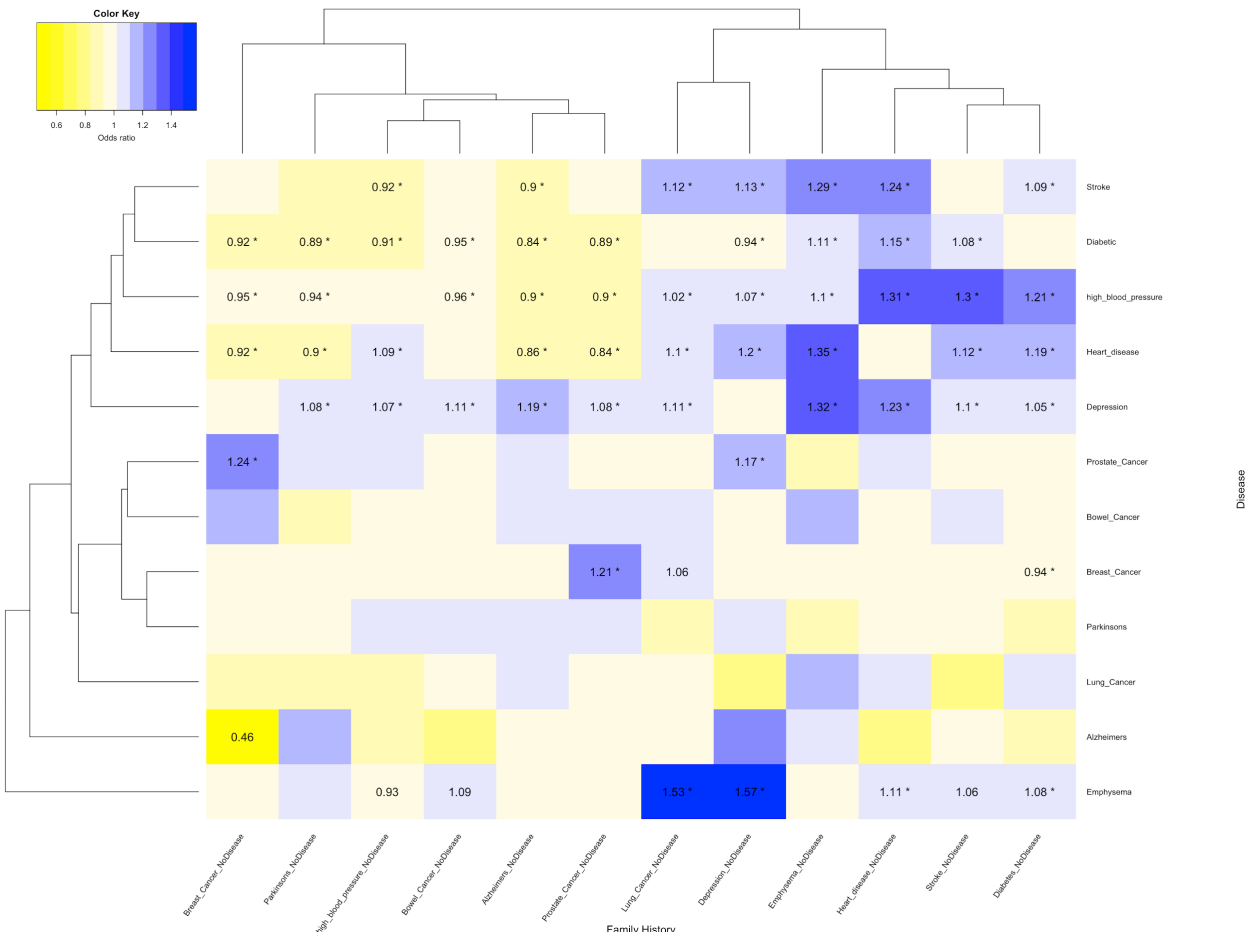
*Paternal indicates paternal family history and no maternal family history. Maternal indicates maternal family history and no paternal family history. Sibling indicates sibling family history and no paternal or maternal family history.
*Sample sizes less than 20 are indicated by "<20".



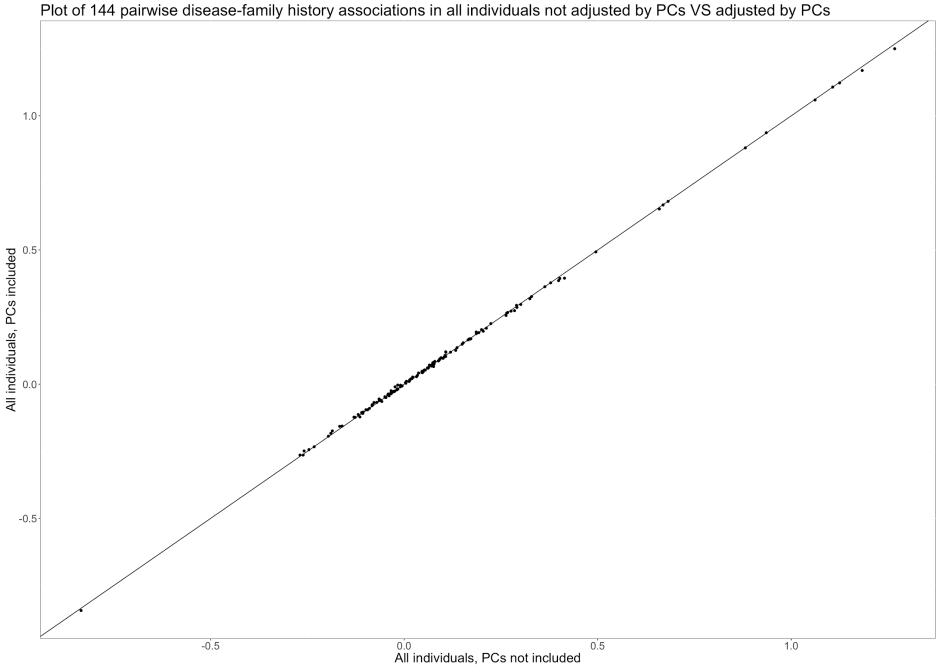
Supplementary Figure 3.1a. All 144 pairwise associations between 12 complex human diseases and their family histories in all individuals, adjusted by age, sex, and 15 principal components. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P*value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age, sex, and 15 principal components.



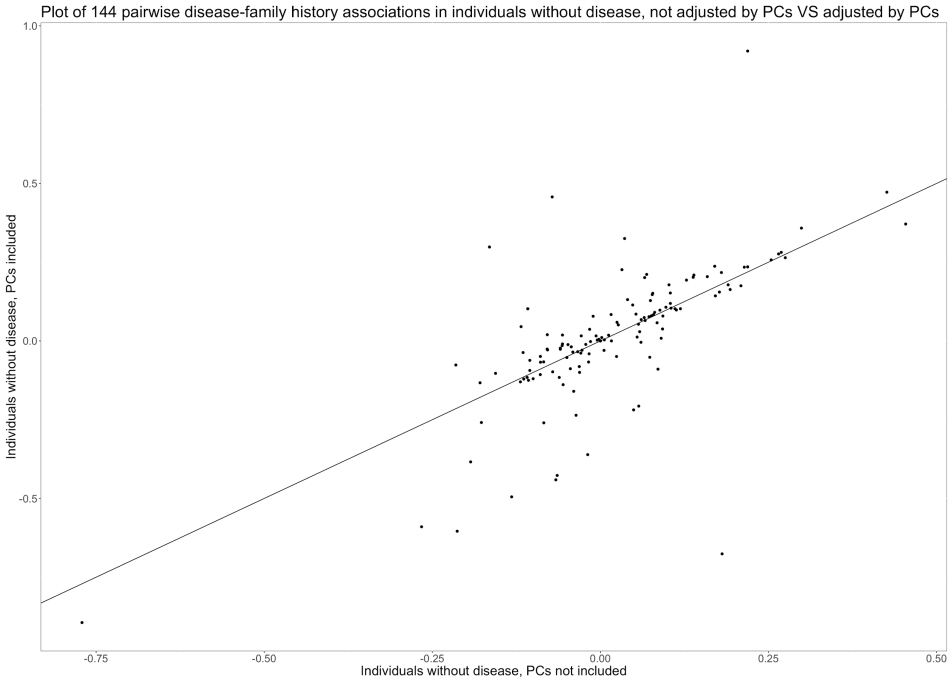
Supplementary Figure 3.1b. All 144 pairwise associations between 12 complex human diseases and their family histories in all individuals, adjusted by age and sex. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P*value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age and sex.



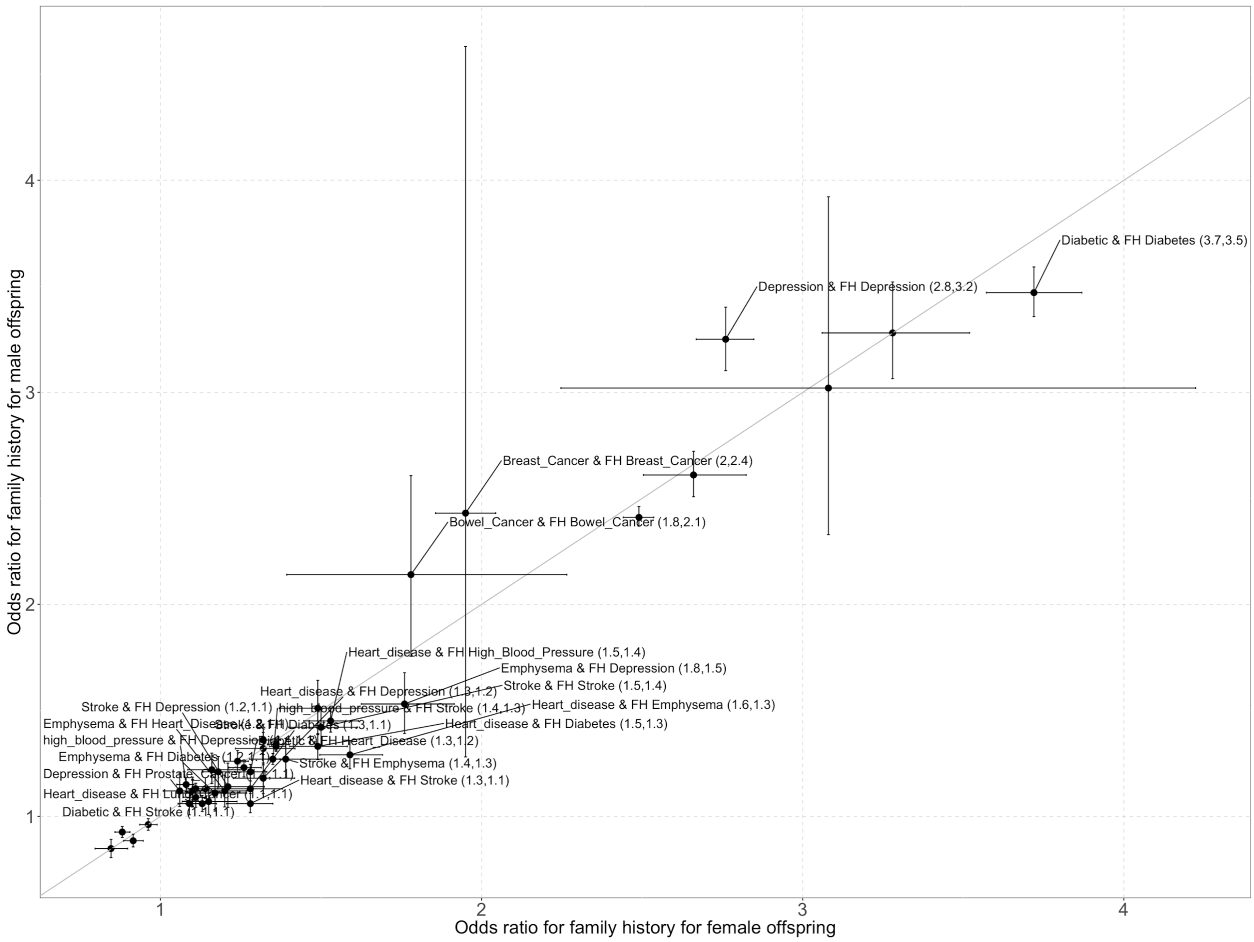
Supplementary Figure 3.2. All 144 pairwise associations between 12 complex human diseases and their family histories for individuals without disease, adjusted by age and sex. Individuals with a positive family history and who had the disease of the family history were removed from analyses. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P*value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age and sex.



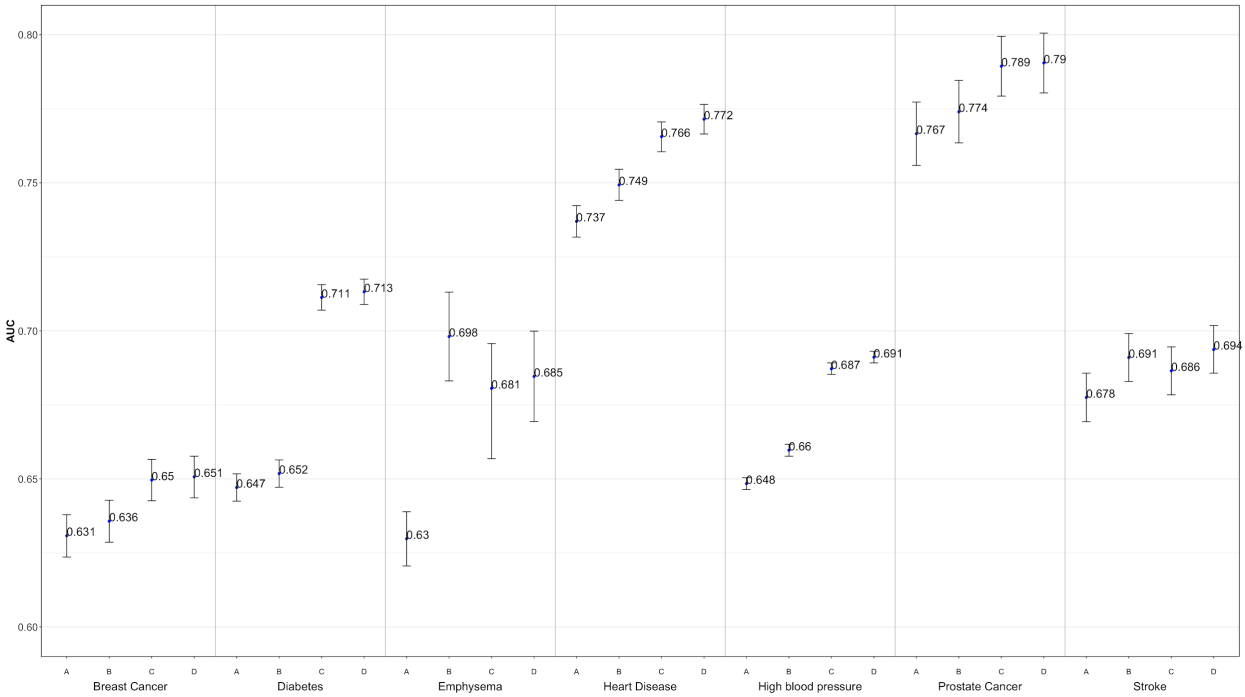
Supplementary Figure 3.3a. Plot of the odds ratio derived from pairwise association between disease and family history. We adjusted by age and sex (x-axis) versus odds ratio derived from pairwise association between disease and family history, adjusted by age, sex, and 15 principal components (y-axis), in all individuals.



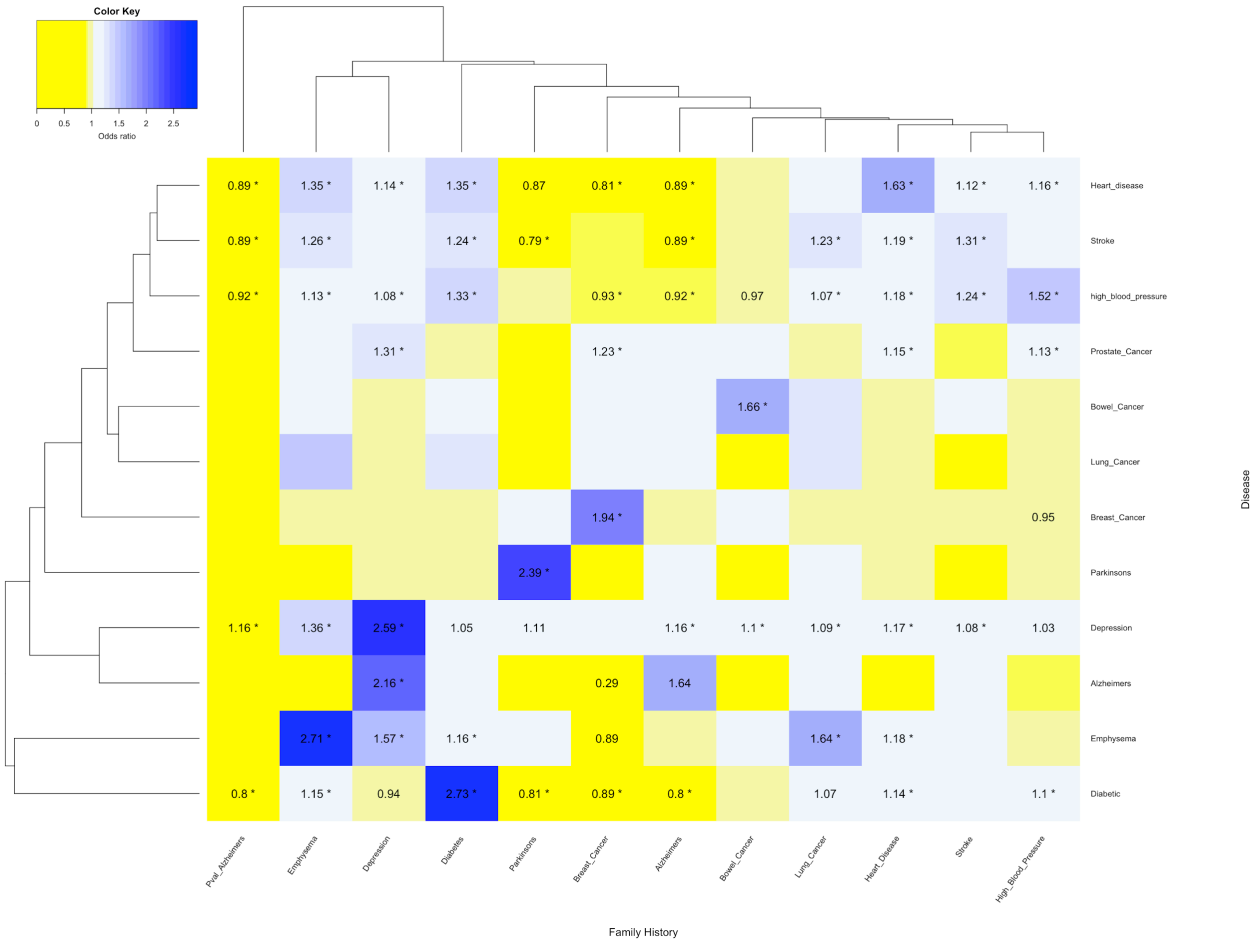
Supplementary Figure 3.3b. Plot of the odds ratio derived from pairwise association between disease and family history. We adjusted by age and sex (x-axis) versus odds ratio derived from pairwise association between disease and family history, adjusted by age, sex, and 15 principal components (y-axis), in individuals without disease.



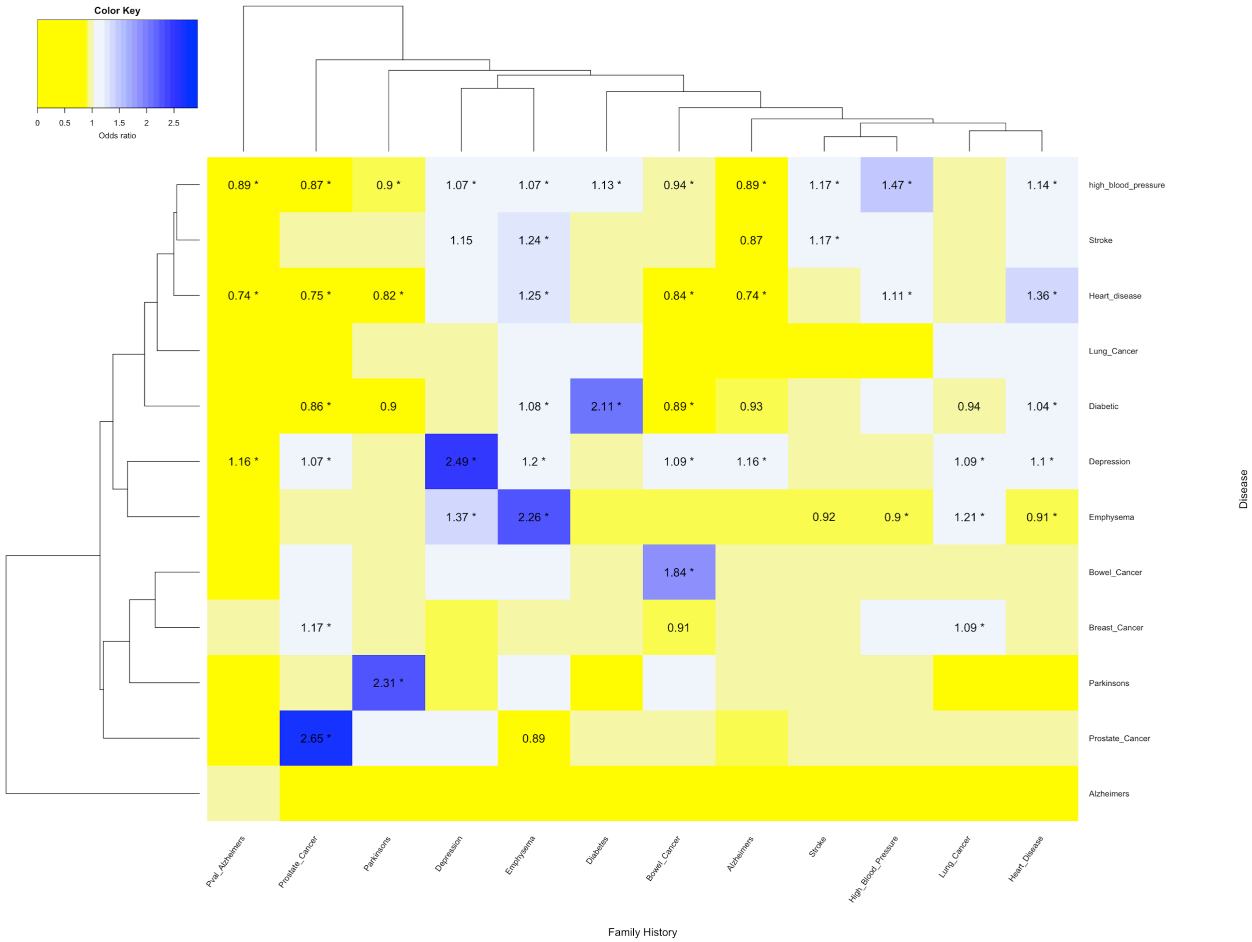
Supplementary Figure 3.4. Comparison of disease-family history associations for female versus male offspring. Disease-family history associations for females (x-axis) are presented against associations for males (y-axis). Horizontal and vertical error bars represent 95% confidence intervals. All points represented are significant at p-value < 0.05 in both analyses.



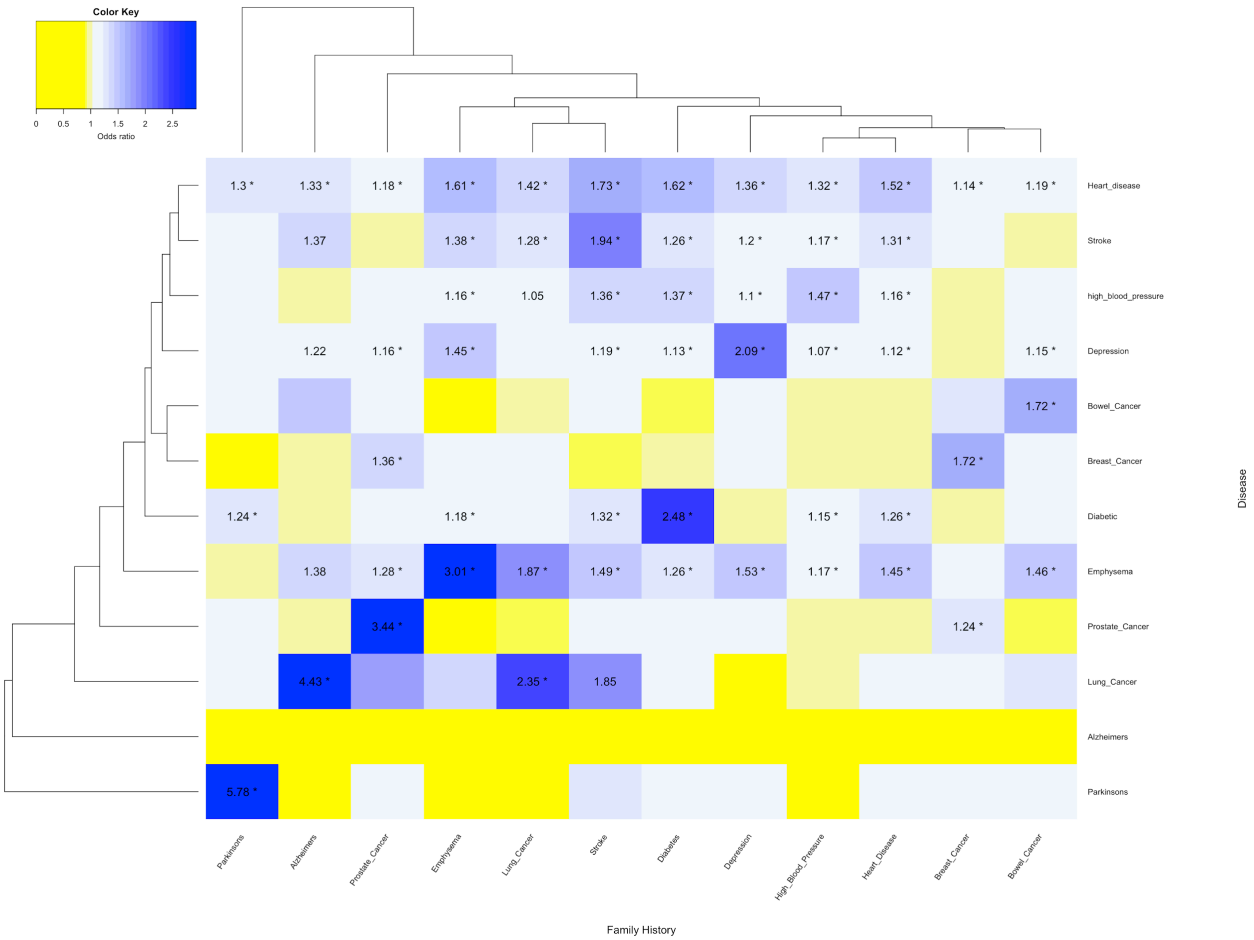
Supplementary Figure 3.5. Comparison of area-under-the-curve (AUC) of 4 different prediction models for the diagnosis of 7 conditions. Model A included age and sex, model B additionally included all family histories identified to be significantly associated at an FDR threshold of 5% with diagnosis of the condition exclusive of the family history of the condition itself, and model D additionally included the family history of the condition itself. Model C included age, sex and family history of the condition only.



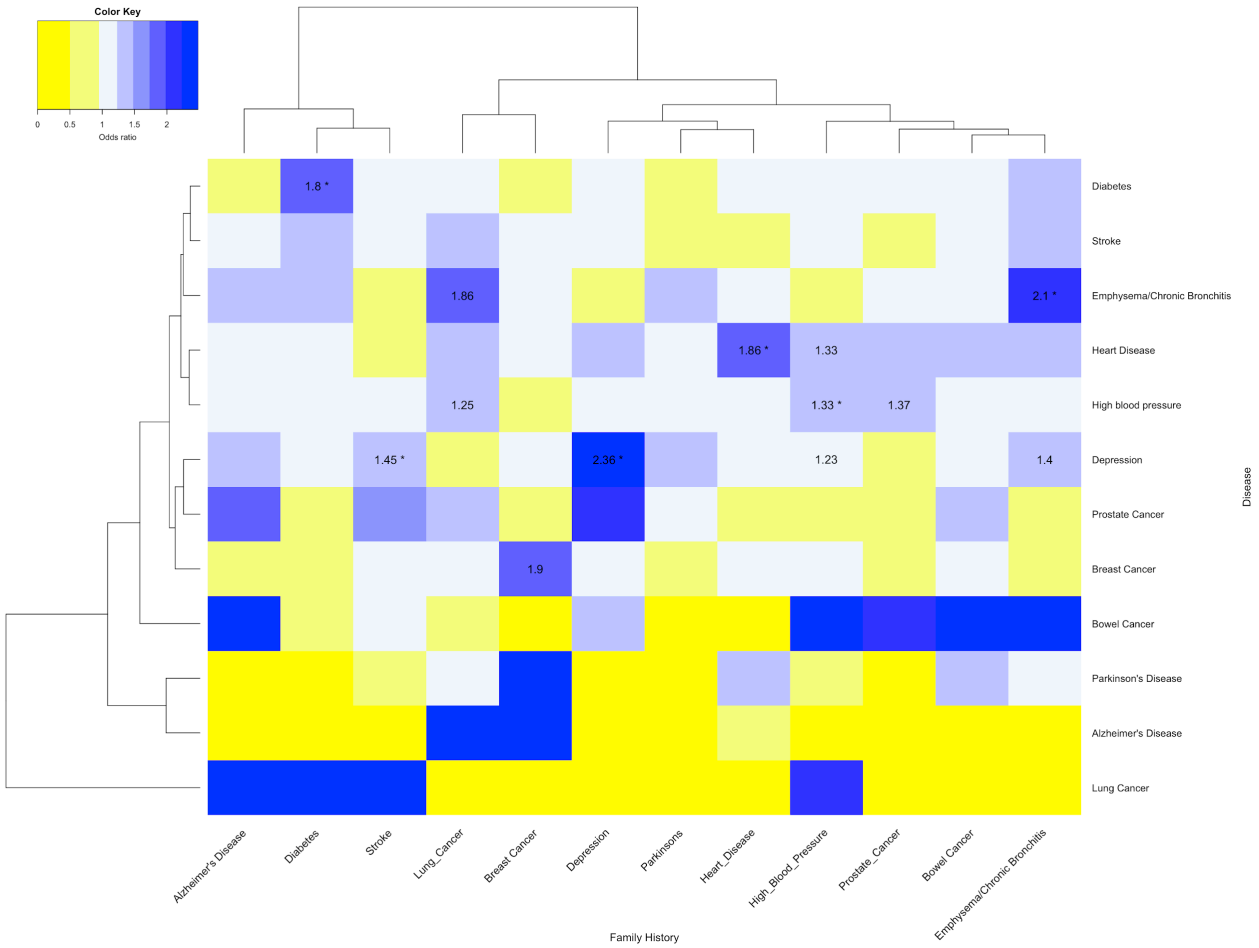
Supplementary Figure 3.6a. Maternal (non-adopted) family history. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P*value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age, sex, and 15 PCs.



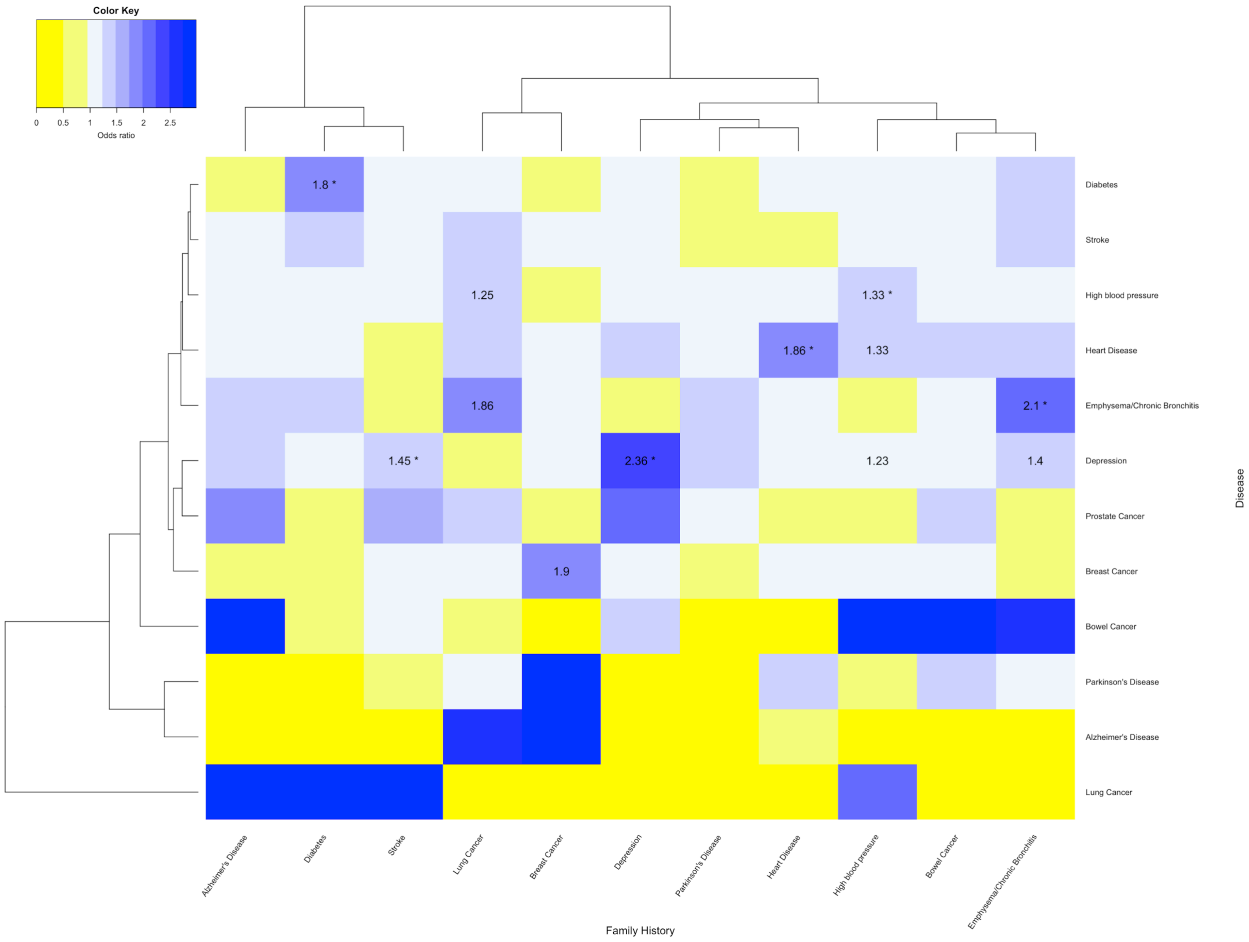
Supplementary Figure 3.6b. Paternal (non-adopted) family history. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P* value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age, sex, and 15 PCs.



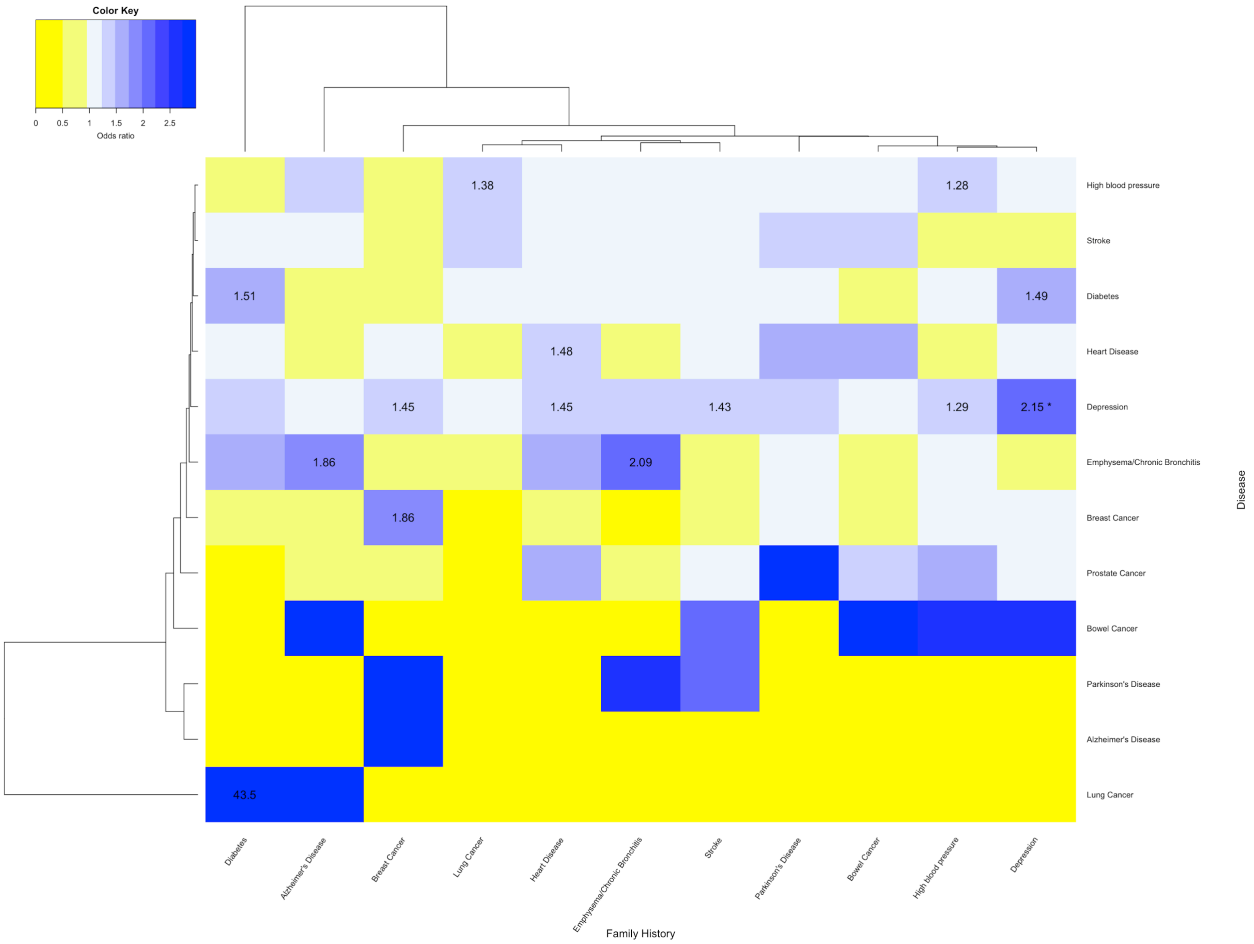
Supplementary Figure 6c. Sibling (non-adopted) family history. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P* value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age, sex, and 15 PCs.



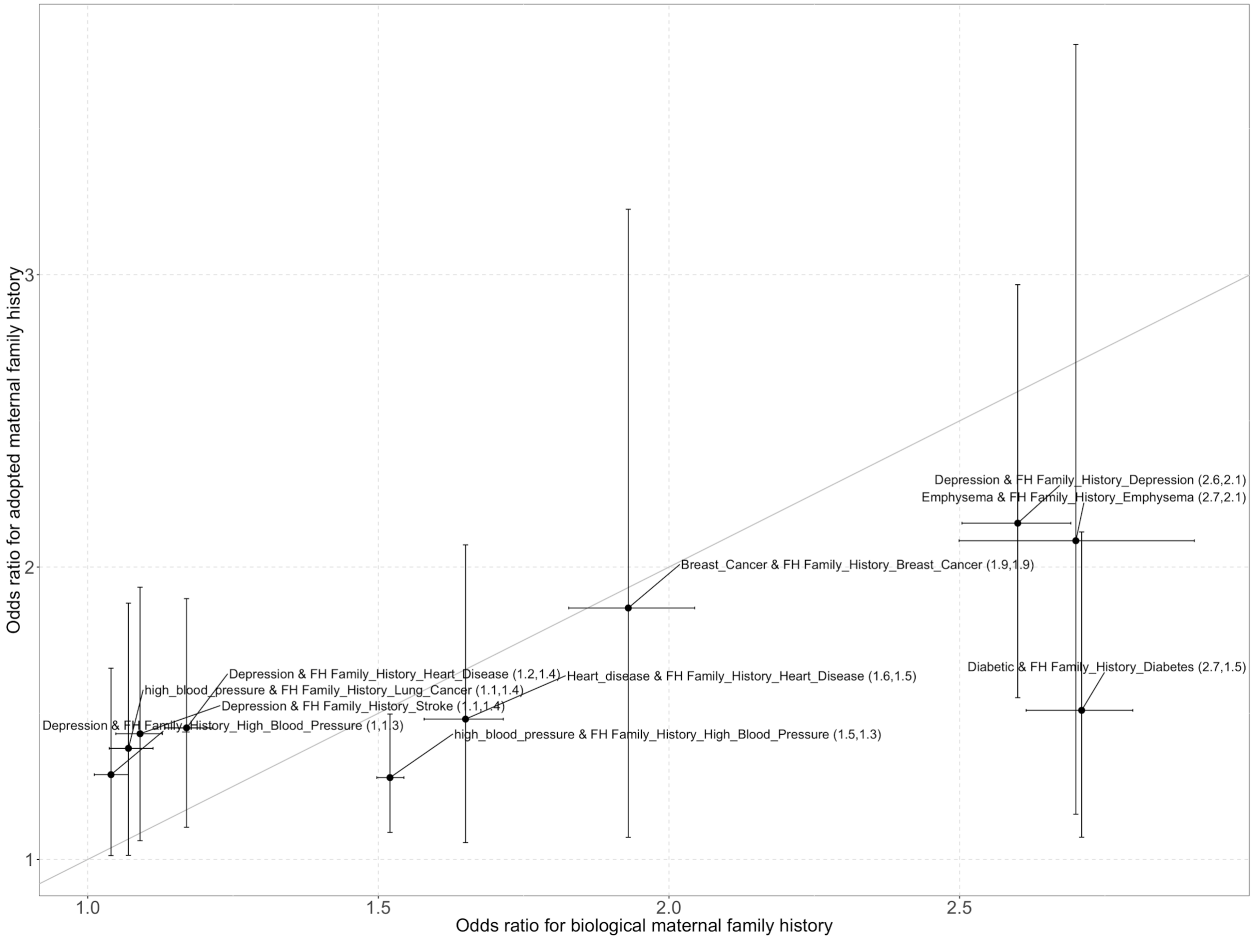
Supplementary Figure 3.7. Adopted cohort findings spanning 144 pairwise associations between 12 complex human diseases and their family histories. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P* value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age, sex, and 15 principal components.



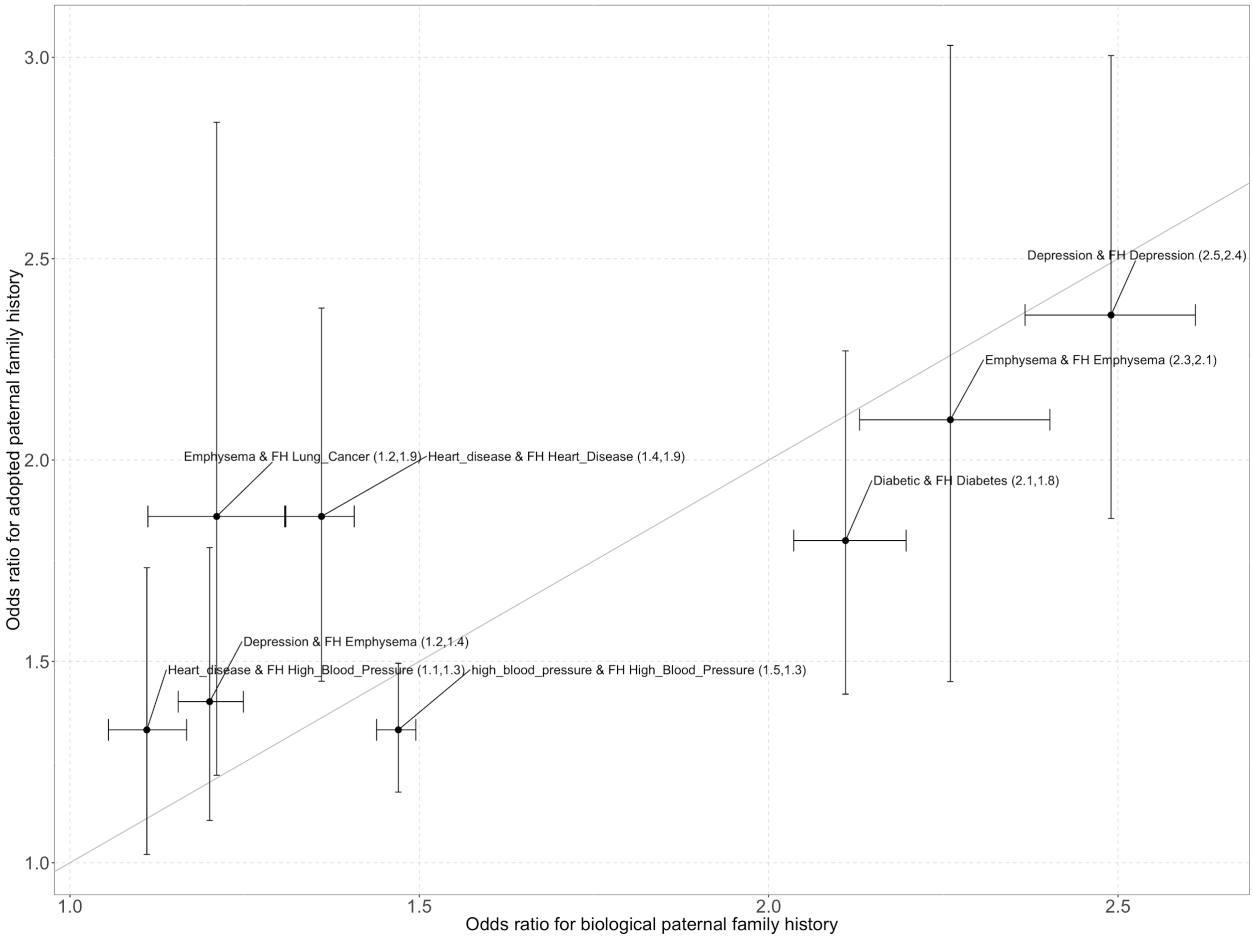
Supplementary Figure 3.8a. Adopted cohort, paternal family history. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P* value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age, sex, and 15 principal components.



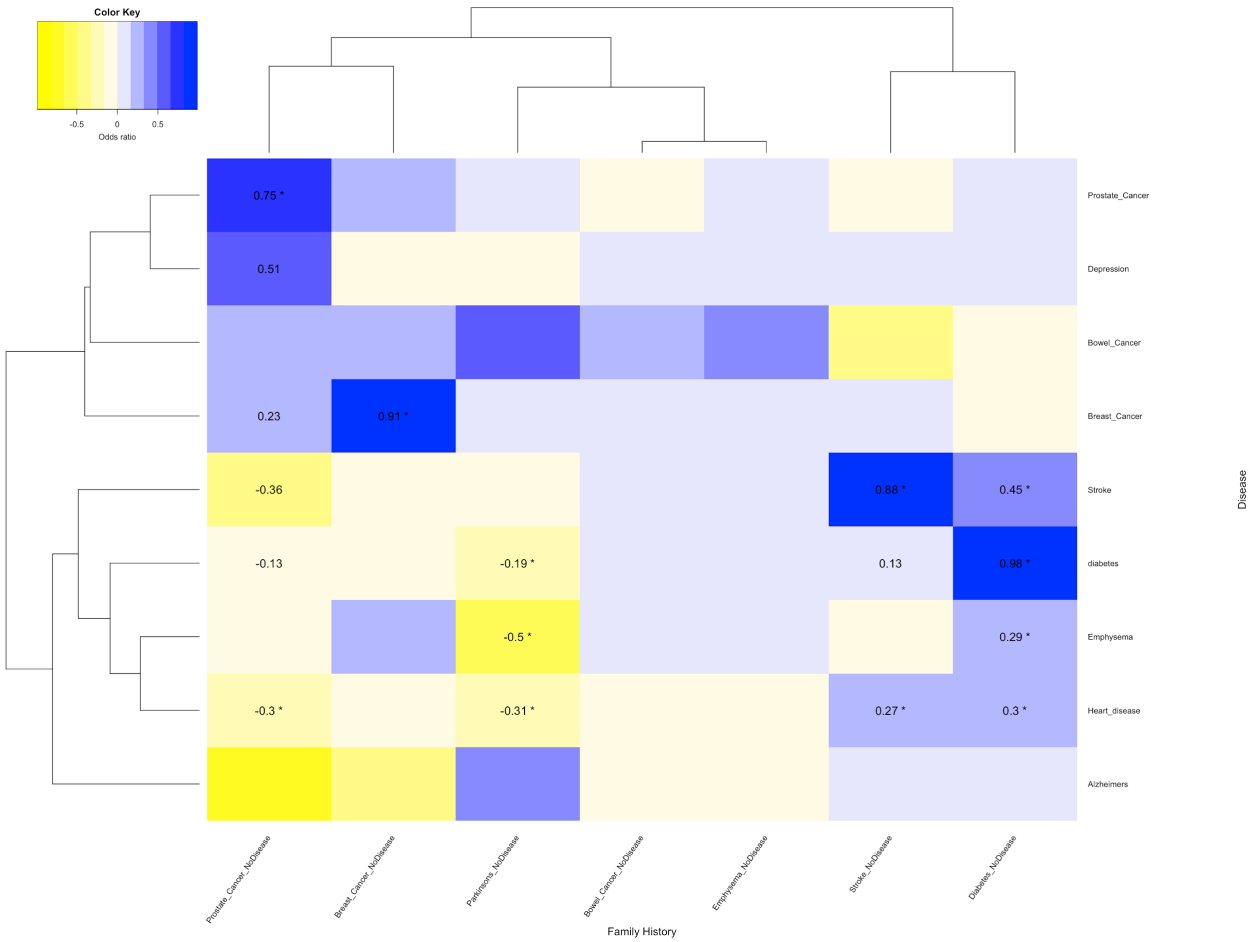
Supplementary Figure 8b. Adopted cohort, maternal family history. Odds ratios are printed inside tiles where the association between family history (presented on x-axis) and disease (y-axis) is significant at a *P* value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, odds ratio greater than 1; yellow, odds ratio less than 1. All associations are adjusted by age, sex, and 15 principal components.



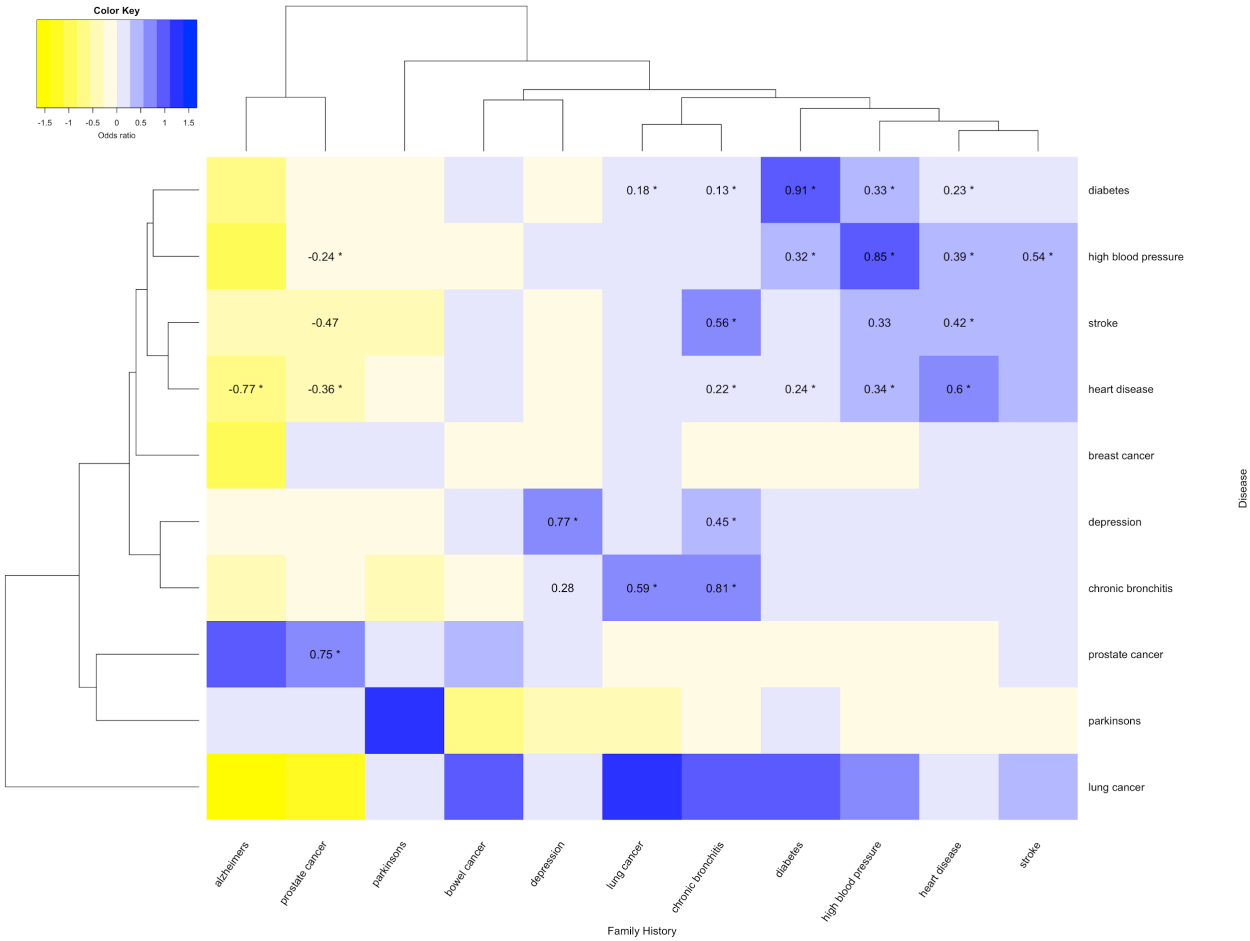
Supplementary Figure 3.9a. Non- adopted maternal compared to adopted maternal family history. Disease-family history associations for non-adopted maternal history (x-axis) are presented against associations for adopted maternal (y-axis). Horizontal and vertical error bars represent 95% confidence intervals. All points represented are significant at p-value < 0.05 in both analyses.



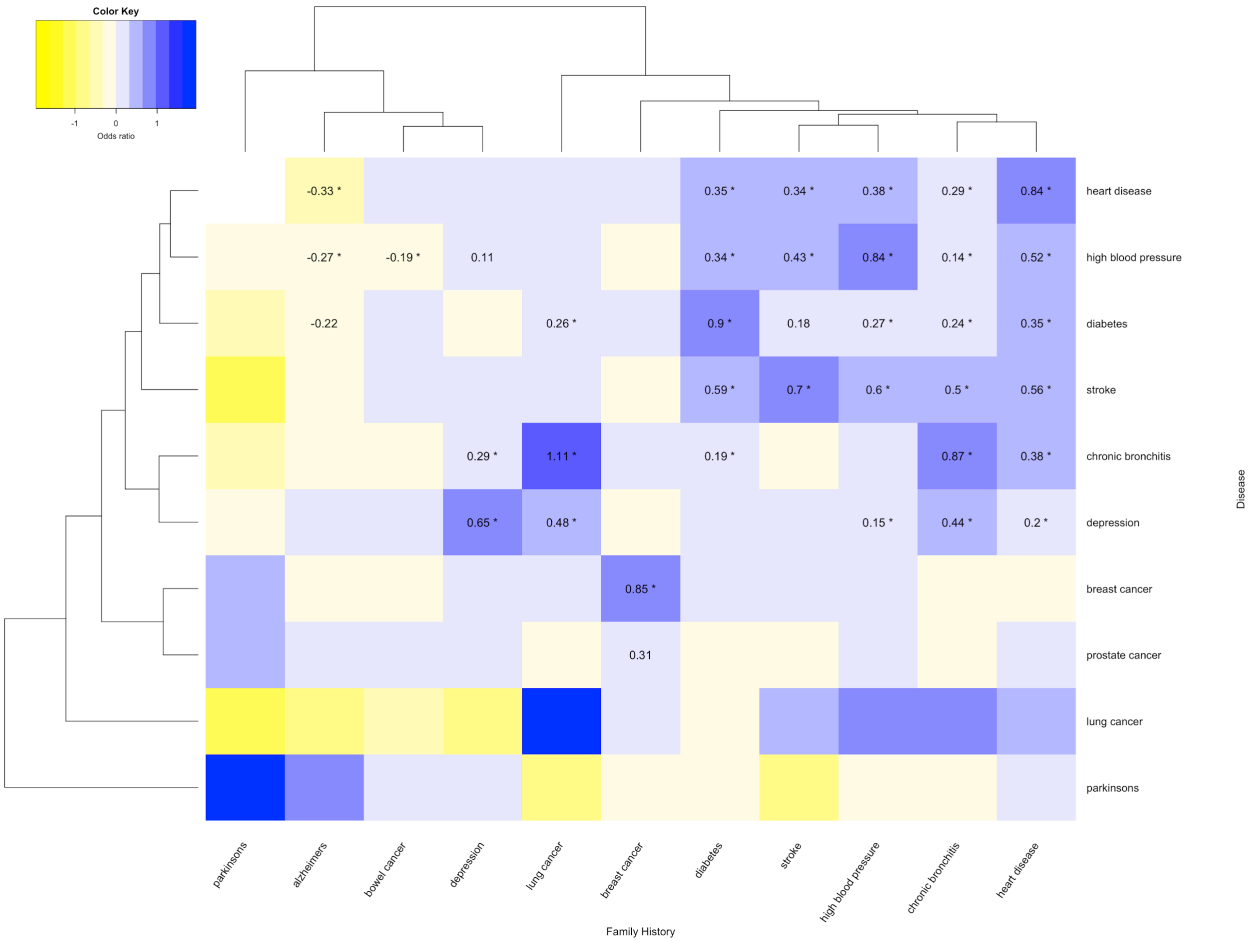
Supplementary Figure 3.9b. Non-adopted paternal compared to adopted paternal history. Disease-family history associations for non-adopted paternal history (x-axis) are presented against associations for adopted paternal (y-axis). Horizontal and vertical error bars represent 95% confidence intervals. All points represented are significant at p-value < 0.05 in both analyses.



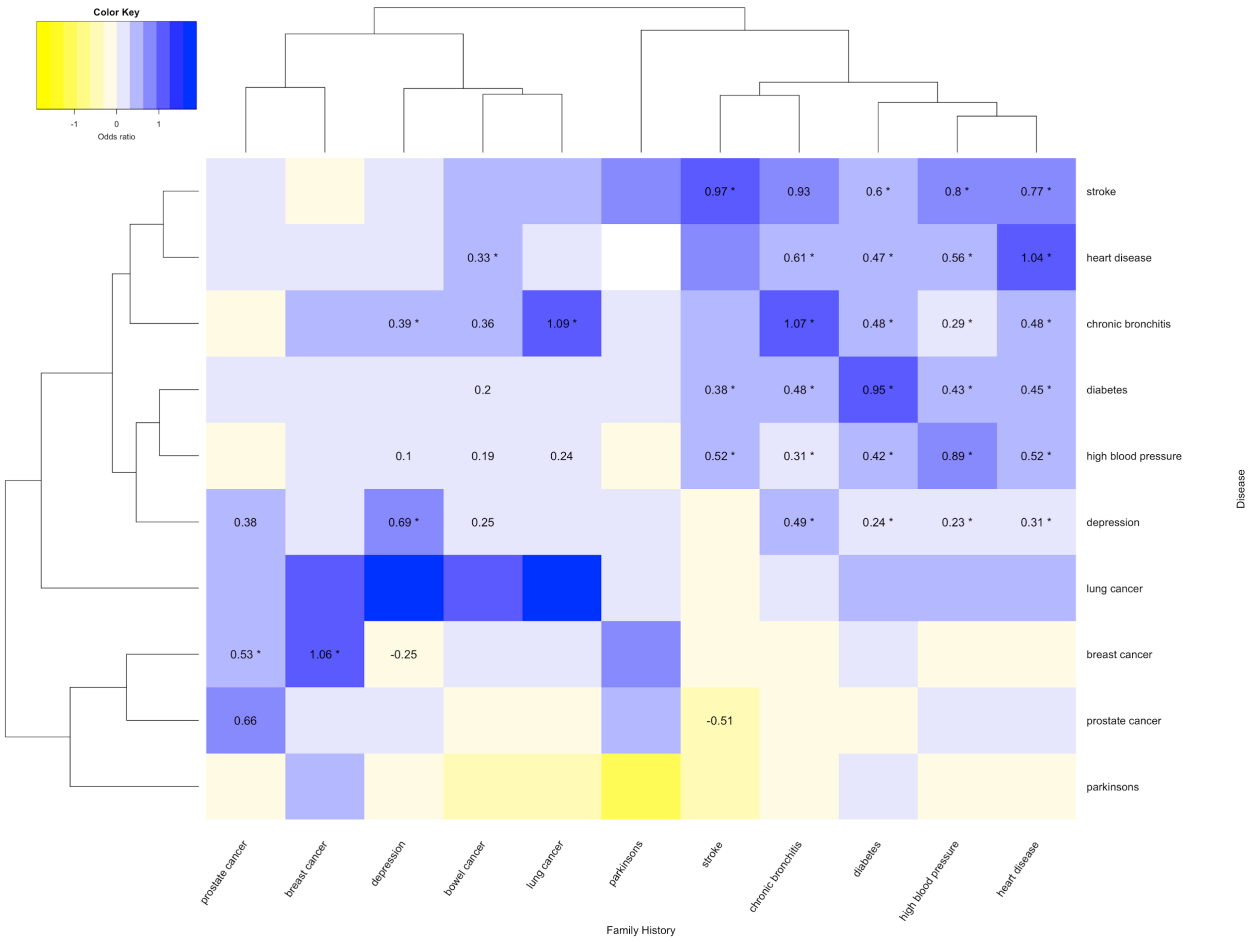
Supplementary Figure 3.10. Genetic correlations between family history of disease and disease in individuals without disease, derived using LD Score Regression. Genetic correlation coefficient (rg) are printed inside tiles where the correlation between family history (presented on x-axis) and disease (y-axis) is significant at a p value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, positive genetic correlation coefficient; yellow, negative genetic correlation coefficient.



Supplementary Figure 3.11a. Genetic correlations between paternal history and self-reported disease, derived using LD Score Regression. Genetic correlation coefficient are printed inside tiles where the correlation between family history (presented on x-axis) and disease (y-axis) is significant at a *P* value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, positive genetic correlation coefficient; yellow, negative genetic correlation coefficient.



Supplementary Figure 3.11b. Genetic correlations between maternal history and self-reported disease, derived using LD Score Regression. Genetic correlation coefficient are printed inside tiles where the correlation between family history (presented on x-axis) and disease (y-axis) is significant at a *P* value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, positive genetic correlation coefficient; yellow, negative genetic correlation coefficient.



Supplementary Figure 3.11c. Genetic correlations between sibling history and self-reported disease, derived using LD Score Regression. Genetic correlation coefficient are printed inside tiles where the correlation between family history (presented on x-axis) and disease (y-axis) is significant at a *P* value less than 0.05. Tiles marked with an asterisk are significant at a false discovery rate (FDR) threshold of 5%. Blue, positive genetic correlation coefficient; yellow, negative genetic correlation coefficient.

Supplementary Table 5.1. ICD-9 diagnosis codes used for ascertaining individuals with T2D.

ICD-9 Codes	Description
250.0	Diabetes mellitus without mention of complication
250.00	Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled
250.02	Diabetes mellitus without mention of complication, type II or unspecified type, uncontrolled
250.1	Diabetes with ketoacidosis
250.10	Diabetes with ketoacidosis, type II or unspecified type, not stated as uncontrolled
250.12	Diabetes with ketoacidosis, type II or unspecified type, uncontrolled
250.2	Diabetes with hyperosmolarity
250.20	Diabetes with hyperosmolarity, type II or unspecified type, not stated as uncontrolled
250.22	Diabetes with hyperosmolarity, type II or unspecified type, uncontrolled
250.3	Diabetes with other coma
250.30	Diabetes with other coma, type II or unspecified type, not stated as uncontrolled
250.32	Diabetes with other coma, type II or unspecified type, uncontrolled
250.4	Diabetes with renal manifestations
250.40	Diabetes with renal manifestations, type II or unspecified type, not stated as uncontrolled
250.42	Diabetes with renal manifestations, type II or unspecified type, uncontrolled
250.5	Diabetes with ophthalmic manifestations
250.50	Diabetes with ophthalmic manifestations, type II or unspecified type, not stated as uncontrolled
250.52	Diabetes with ophthalmic manifestations, type II or unspecified type, uncontrolled
250.6	Diabetes with neurological manifestations
250.60	Diabetes with neurological manifestations, type II or unspecified type, not stated as uncontrolled

250.62	Diabetes with neurological manifestations, type II or unspecified type, uncontrolled
250.7	Diabetes with peripheral circulatory disorders
250.70	Diabetes with peripheral circulatory disorders manifestations, type II or unspecified type, not stated as uncontrolled
250.72	Diabetes with peripheral circulatory disorders manifestations, type II or unspecified type, uncontrolled
250.8	Diabetes with other specified manifestations
250.80	Diabetes with other specified manifestations, type II or unspecified type, not stated as uncontrolled
250.82	Diabetes with other specified manifestations, type II or unspecified type, uncontrolled
250.9	Diabetes with unspecified complication
250.90	Diabetes with unspecified complication, type II or unspecified type, not stated as uncontrolled
250.92	Diabetes with unspecified complication, type II or unspecified type, uncontrolled

Supplementary Table 5.2. The list below of T2D medications and insulin therapy by drug name and brand names was used in our filtration criteria.

Drug name	Brand name
Metformin/ Biguanides	Glucophage Glumetza
Glyburide	Diabeta, Glycron, Glynase, Micronase
Glipizide	Glucotrol, Glucotrol XL
Glimepiride	Amaryl
Repaglinide (Meglitinides)	Prandin
Nateglinide	Starlix
Rosiglitazone, Pioglitazone (Thiazolidinediones)	Avandia, Actos
Sitagliptin	Januvia
Saxagliptin	Onglyza
Linagliptin	Tradjenta
Dapagliflozin	Farxiga
Alpha-glucosidase inhibitors	Acarbose (Precose) Miglitol (Glycet)

Drug name	Brand name
Liraglutide	Victoza
Canagliflozin	Invokana
Dapagliflozin	Farxiga
Insulin glulisine	Apidra
Insulin lispro	Humalog
Insulin aspart	Novolog
Insulin glargine	Lantus
Insulin detemir	Levemir
Insulin isophane	Humulin N Novolin N
DPP 4 inhibitors	Galvus (Vildagliptin)
Exenatide	Byetta Bydureon
Sulfonylureas	Diabinese, Tolinase (Tolazamide) Tolbutamide (Orinase) Acetohexamide (Dymelor)

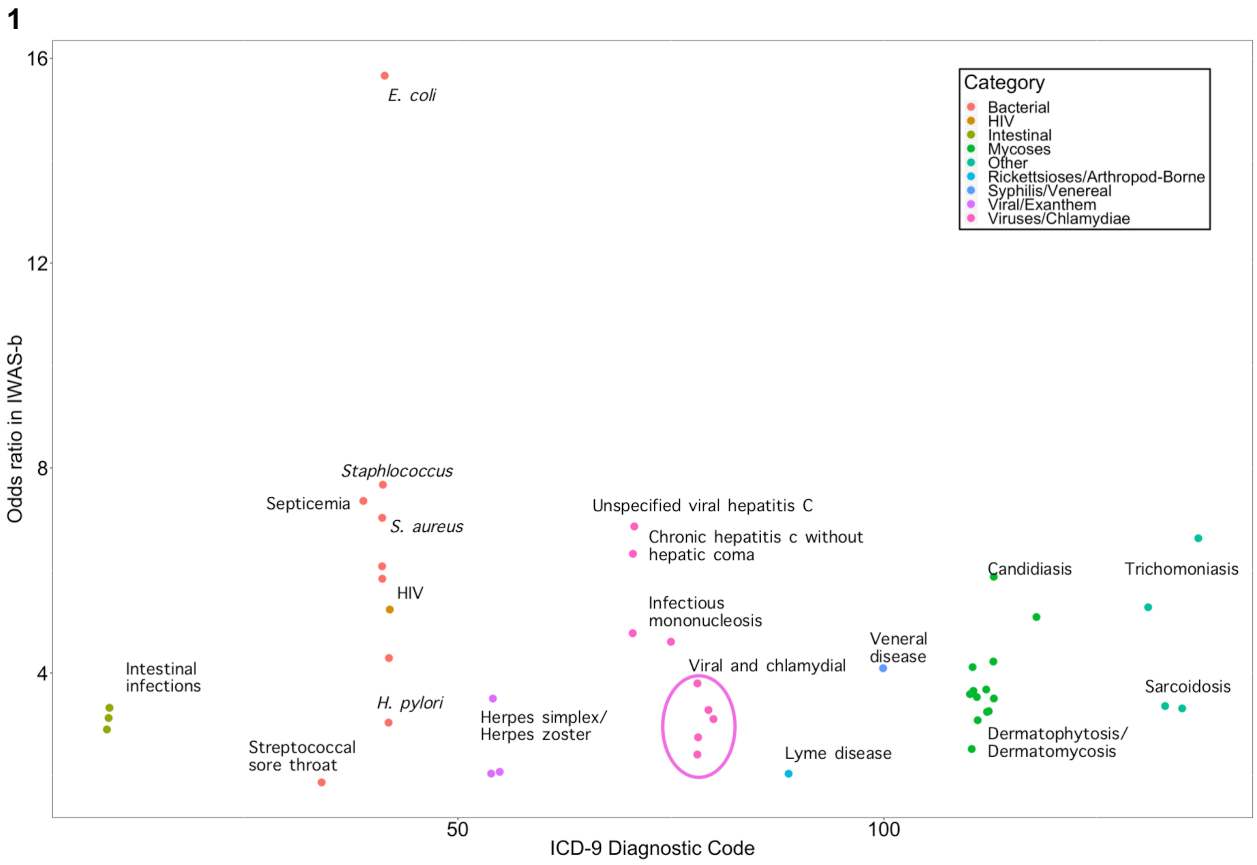
Supplementary Table 5.3. ICD-9 diagnosis codes used for ascertaining individuals who are overweight/obese.

ICD-9 Codes	Description
278	Overweight, obesity, and other hyperalimentation
278.0	Overweight and obesity
278.00	Obesity, unspecified
278.01	Morbid obesity
278.02	Overweight
278.03	Obesity hypoventilation syndrome
V85.2	Body Mass Index between 25-29, adult
V85.21	Body Mass Index 25.0-25.9, adult
V85.22	Body Mass Index 26.0-26.9, adult
V85.23	Body Mass Index 27.0-27.9, adult
V85.24	Body Mass Index 28.0-28.9, adult
V85.25	Body Mass Index 29.0-29.9, adult
V85.30	Body Mass Index between 30-39, adult
V85.31	Body Mass Index 30.0-30.9, adult
V85.32	Body Mass Index 31.0-31.9, adult
V85.33	Body Mass Index 32.0-32.9, adult
V85.34	Body Mass Index 33.0-33.9, adult
V85.35	Body Mass Index 34.0-34.9, adult
V85.36	Body Mass Index 35.0-35.9, adult
V85.37	Body Mass Index 36.0-36.9, adult
V85.38	Body Mass Index 37.0-37.9, adult
V85.39	Body Mass Index 38.0-38.9, adult

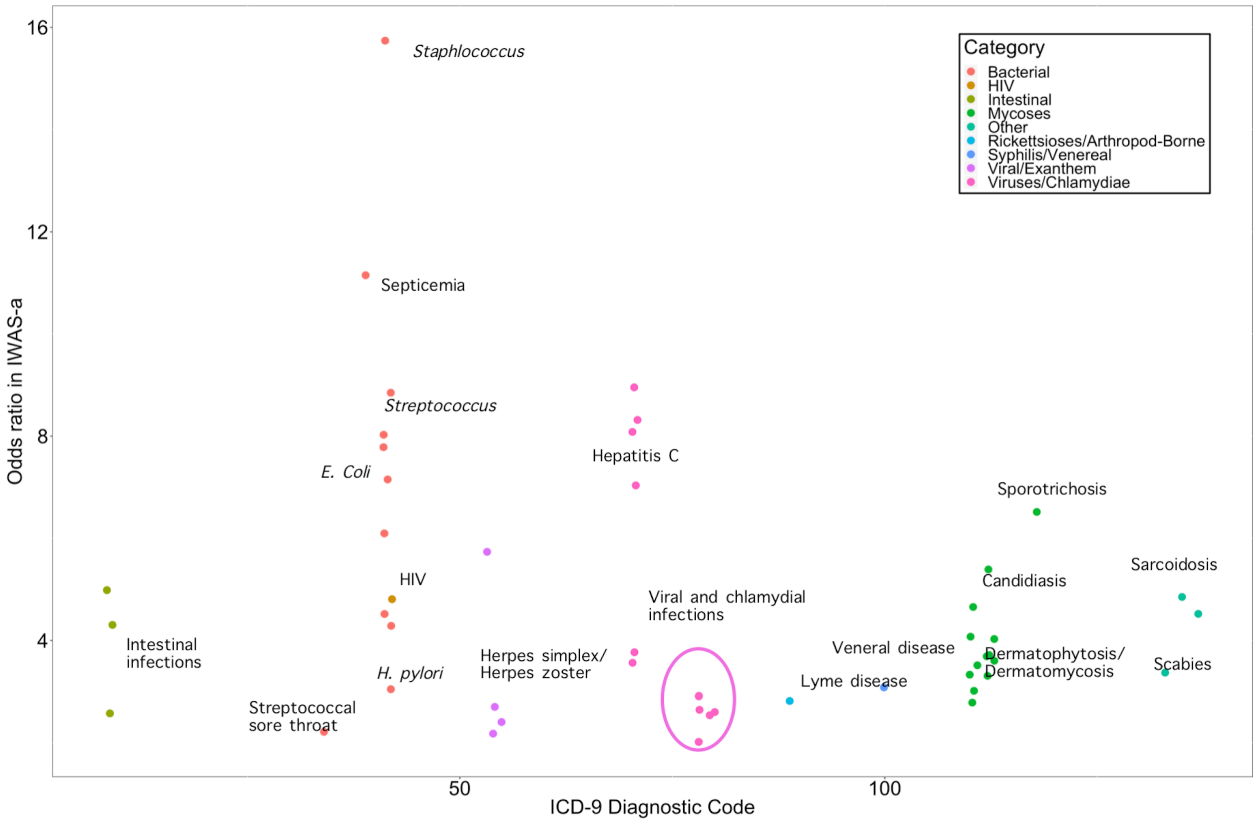
V85.4	Body Mass Index 40 and over, adult

Supplementary Table 5.4. Summary of the categories of infectious diseases, the range of ICD-9 codes classified for each category, and the number of ICD-9 codes included in our study from each category in IWAS-b and IWAS-a, respectively.

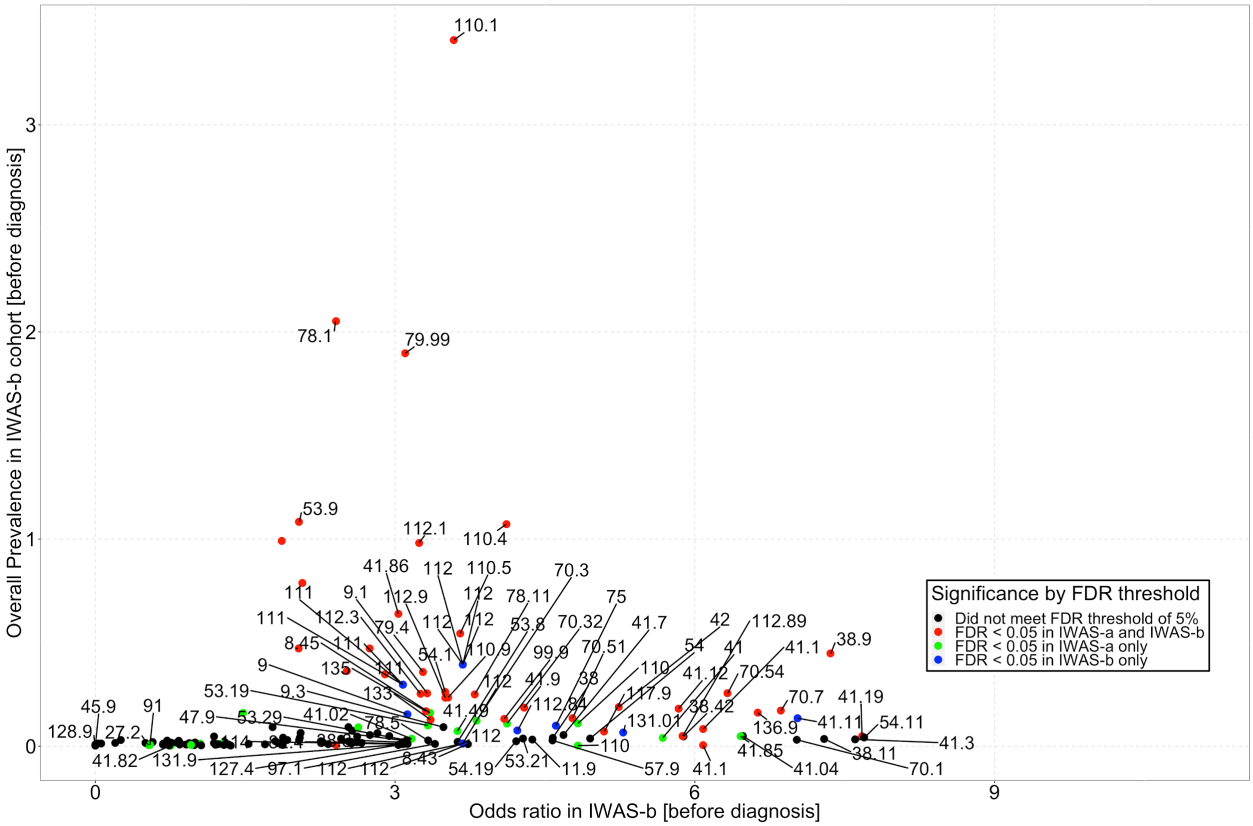
Category	**ICD-9 range of category**	**Number of ICD-9 codes in IWAS-b**	**Number of ICD-9 codes in IWAS-a**
Intestinal infectious diseases	001-009	20	24
Tuberculosis	010-018	3	2
Zoonotic bacterial diseases	020-027	1	0
Other bacterial diseases	030-041	47	57
Human immunodeficiency virus (HIV) infection	042-044	1	1
Poliomyelitis and other non-arthropod-borne viral diseases of the central nervous system	045-049	4	4
Viral diseases accompanied by exanthem	050-059	37	38
Arthropod-borne viral diseases	060-069	3	4
Other diseases due to viruses and chlamydiae	070-079	46	47
Rickettsioses and other arthropod-borne diseases	080-089	9	11
Syphilis and other venereal diseases	090-099	21	21
Other spirochetal diseases	100-104	3	3
Mycoses	110-118	33	36
Helminthiases	120-129	7	9
Other infectious and parasitic diseases	130-136	15	15
Late effects of infectious and parasitic diseases	137-139	2	2
Total Number of ICDs		252	274



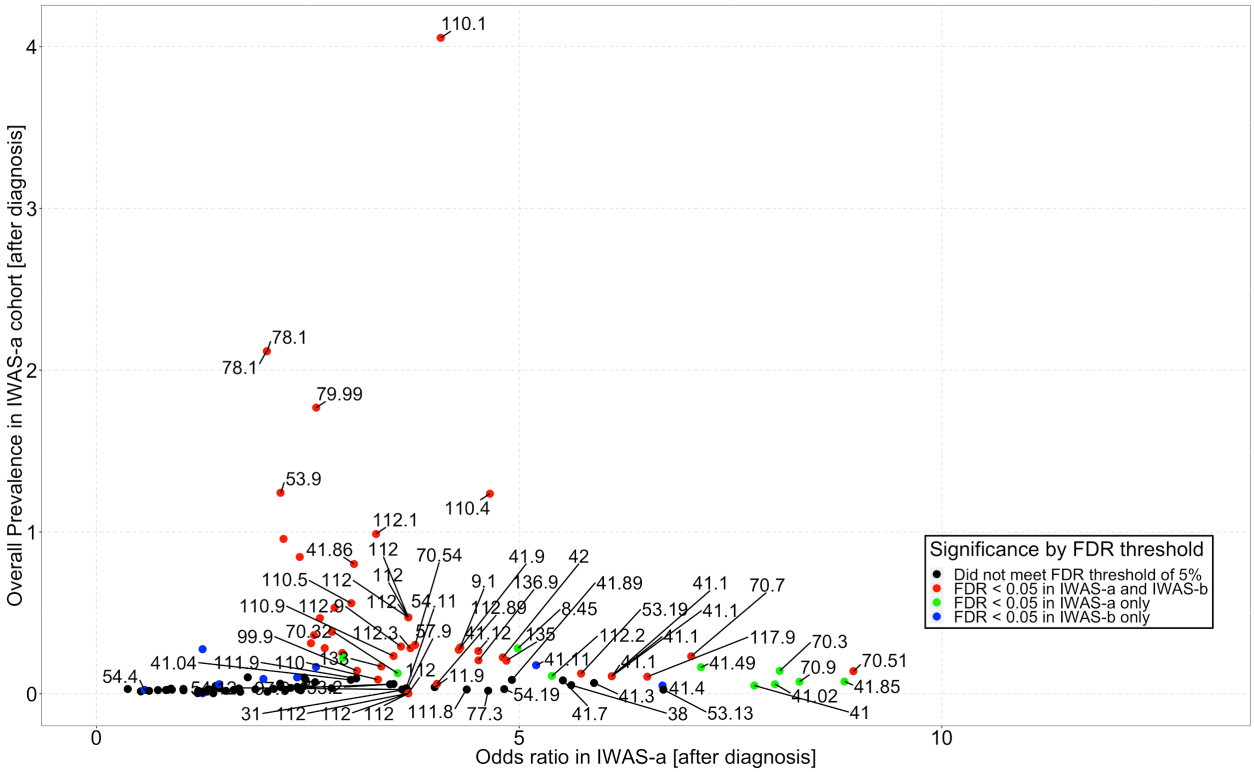
Supplementary Figure 5.2a. Plot of ICD-9 diagnostic codes and corresponding odds ratio identified from IWAS-b. All displayed points achieved significance at an FDR threshold of 5% in both the training and validation sets.



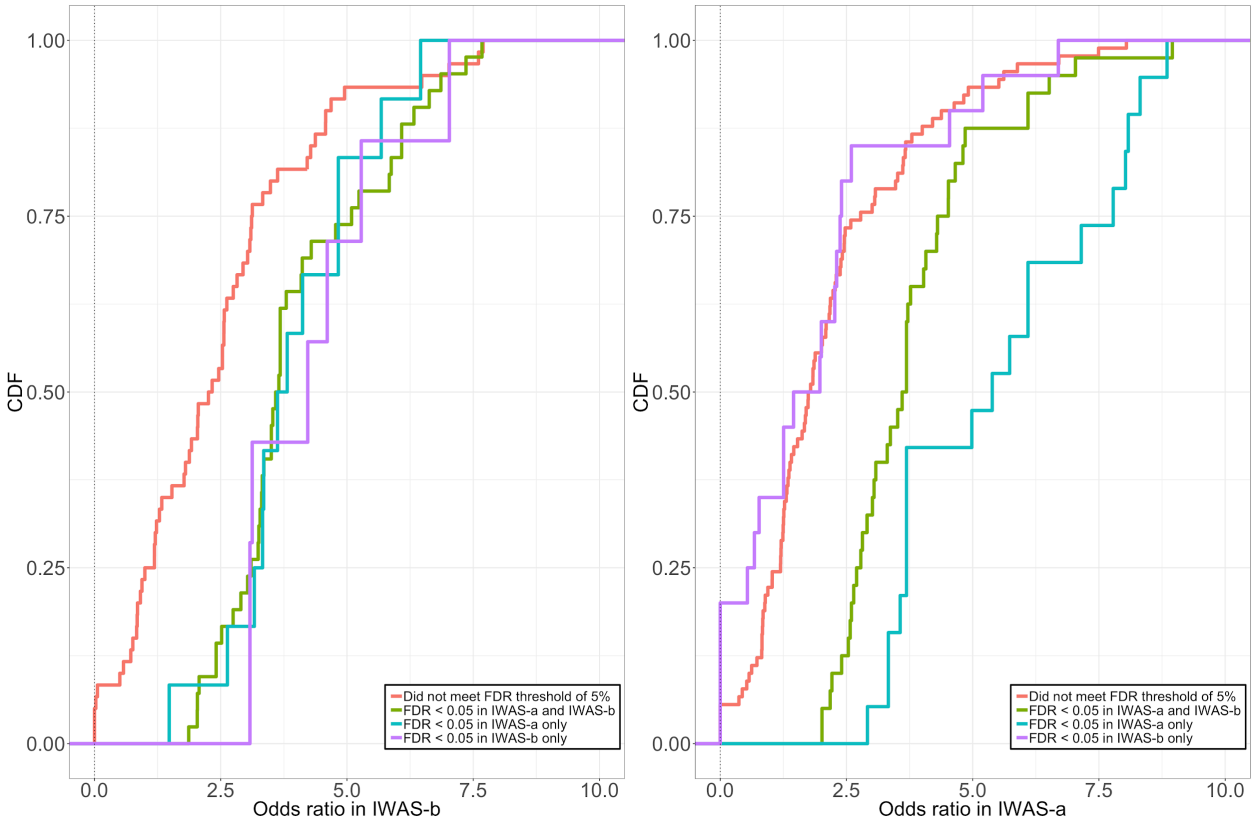
Supplementary Figure 5.2b. Plot of ICD-9 diagnostic codes and corresponding odds ratio identified from IWAS-a. All displayed points achieved significance at an FDR threshold of 5% in both the training and validation sets.



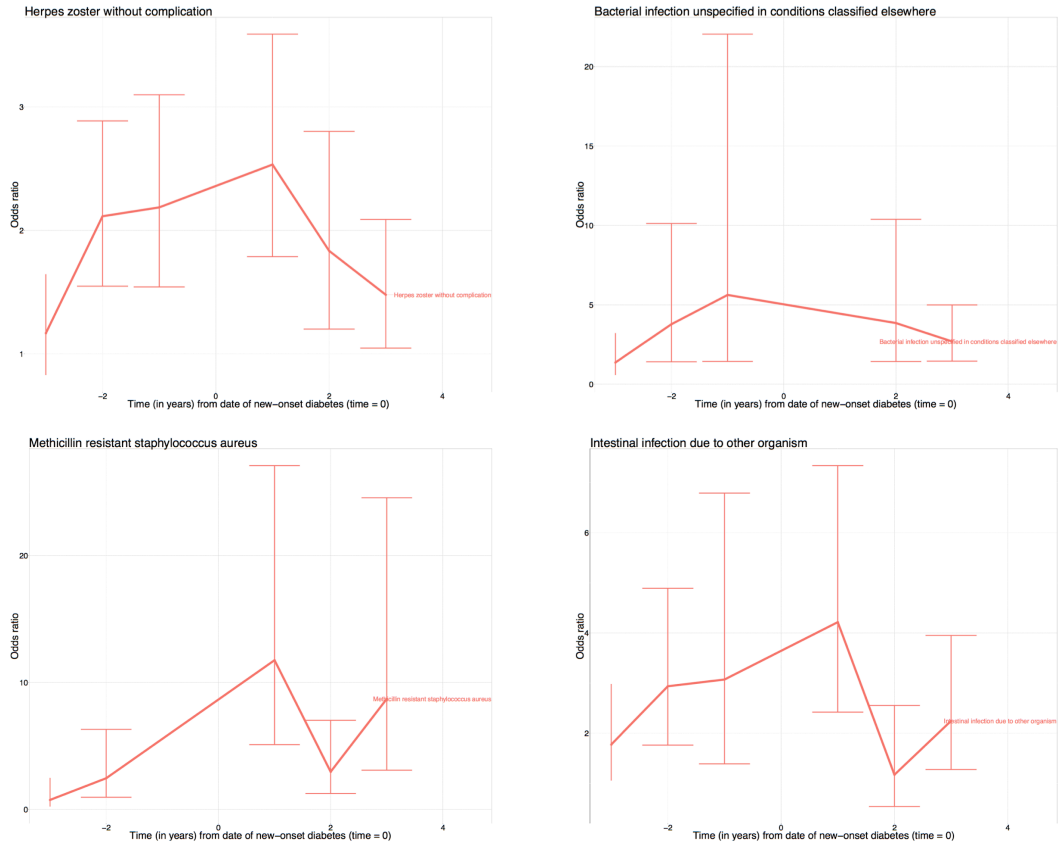
Supplementary Figure 5.3A. Plot of odds ratio for each ICD-9 diagnostic code from IWAS-b analysis versus prevalence of the ICD-9 diagnostic code within IWAS-b case and control groups. Prevalence is calculated as the total number of patients with the ICD-9 diagnostic code in the 24-month infection window before T2D diagnosis divided by the total number of individuals in the IWAS-b case and control groups. Each point is labeled with the ICD-9 diagnostic code.



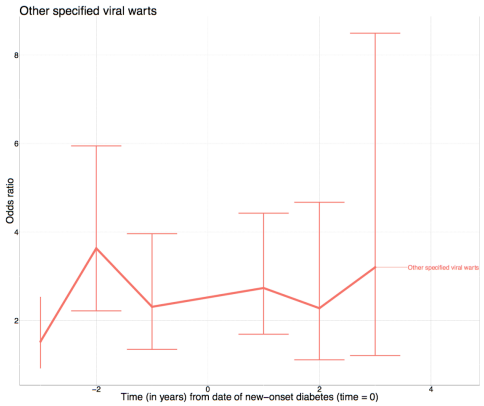
Supplementary Figure 5.3B. Plot of odds ratio for each ICD-9 diagnostic code from IWAS-a analysis versus prevalence of the ICD-9 diagnostic code within IWAS-a case and control groups. Prevalence is calculated as the total number of patients with the ICD-9 diagnostic code in the 24-month infection window after T2D diagnosis divided by the total number of individuals in our IWAS-a case and control groups. Each point is labeled with the ICD-9 diagnostic code.

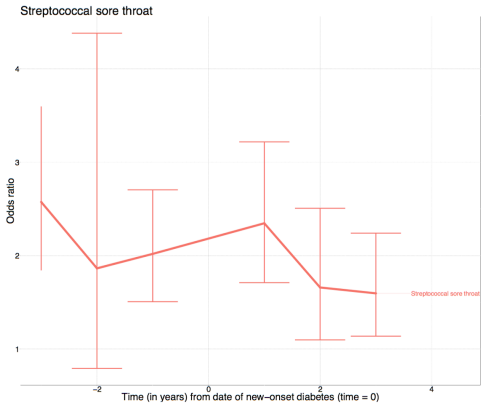


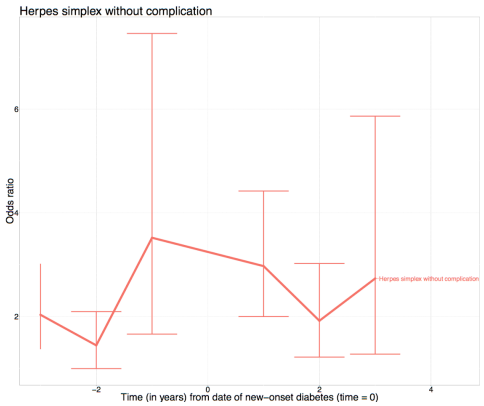
Supplementary Figure 5.4. CDF plot of distribution of effect sizes (top) in IWAS-b and IWAS-a.

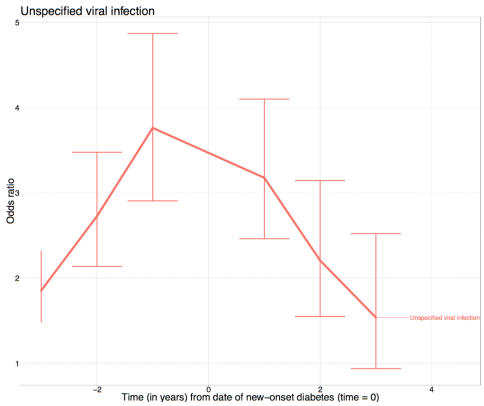


Supplementary Figure 5.5. Odds ratios for infections identified at varying 1-year time intervals within a three-year period before and after the date of documented T2D. *Figure continues on pages following.*


Other specified viral warts
Odds ratio
Time (in years) from date of new−onset diabetes (time = 0)
Other specified viral warts


Streptococcal sore throat
Odds ratio
Time (in years) from date of new−onset diabetes (time = 0)
Streptococcal sore throat


Herpes simplex without complication
Odds ratio
Time (in years) from date of new−onset diabetes (time = 0)
Herpes simplex without complication


Unspecified viral infection
Odds ratio
Time (in years) from date of new−onset diabetes (time = 0)
Unspecified viral infection

References

1. Guttmacher AE, Collins FS, Carmona RH. The family history--more important than ever. N Engl J Med. 2004;351: 2333–2336.

2. Valdez R, Yoon PW, Qureshi N, Green RF, Khoury MJ. Family history in public health practice: a genomic tool for disease prevention and health promotion. Annu Rev Public Health. 2010;31: 69–87 1 p following 87.

3. Rasooly D, Ioannidis JPA, Khoury MJ, Patel CJ. Family History–Wide Association Study to Identify Clinical and Environmental Risk Factors for Common Chronic Diseases. American Journal of Epidemiology. 2019. pp. 1563–1568. doi:10.1093/aje/kwz125

4. Denny JC, Ritchie MD, Basford MA, Pulley - … JM, 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. academic.oup.com. 2010. Available: https://academic.oup.com/bioinformatics/article-abstract/26/9/1205/201211

5. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101: 5–22.

6. Centers for Disease Control and Prevention: National Center for Health Statistics. NHANES - National Health and Nutrition Examination Survey Homepage. [cited 2 Dec 2016]. Available: https://www.cdc.gov/nchs/nhanes/index.htm

7. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2011;34 Suppl 1: S62–9.

8. Vogelmeier CF, Criner GJ, Martinez FJ, Others. Global Initiative for Chronic Obstructive Lung Disease Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease 2017 Report: GOLD executive summary. Am J Respir Crit Care Med. 2017;195: 557–582.

9. Lumley T. Analysis of Complex Survey Samples. J Stat Softw. 2004;9: 1–19.

10. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995. pp. 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

11. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. 2010. pp. 1–48. Available: http://www.jstatsoft.org/v36/i03/

12. Meigs JB, Cupples LA, Wilson PW. Parental transmission of type 2 diabetes: the

Framingham Offspring Study. Diabetes. 2000;49: 2201–2207.

13. Harrison TA, Hindorff LA, Kim H, Wines RCM, Bowen DJ, McGrath BB, et al. Family history of diabetes as a potential public health tool. Am J Prev Med. 2003;24: 152–159.

14. Hariri S, Yoon PW, Qureshi N, Valdez R, Scheuner MT, Khoury MJ. Family history of type 2 diabetes: a population-based screening tool for prevention? Genet Med. 2006;8: 102–108.

15. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47: 1236–1241.

16. Eliassen B-M, Melhus M, Tell GS, Borch KB, Braaten T, Broderstad AR, et al. Validity of self-reported myocardial infarction and stroke in regions with Sami and Norwegian populations: the SAMINOR 1 Survey and the CVDNOR project. BMJ Open. 2016;6: e012717.

17. Oksanen T, Kivimäki M, Pentti J, Virtanen M, Klaukka T, Vahtera J. Self-report as an indicator of incident disease. Ann Epidemiol. 2010;20: 547–554.

18. Janssens ACJW, Henneman L, Detmar SB, Khoury MJ, Steyerberg EW, Eijkemans MJC, et al. Accuracy of self-reported family history is strongly influenced by the accuracy of self-reported personal health status of relatives. J Clin Epidemiol. 2012;65: 82–89.

19. Bensen JT, Liese AD, Rushing JT, Province M, Folsom AR, Rich SS, et al. Accuracy of proband reported family history: the NHLBI Family Heart Study (FHS). Genet Epidemiol. 1999;17: 141–150.

20. Liu JZ, Erlich Y, Pickrell JK. Case–control association mapping by proxy using family history of disease. Nat Genet. 2017;49: 325.

21. InterAct Consortium, Scott RA, Langenberg C, Sharp SJ, Franks PW, Rolandsson O, et al. The link between family history and risk of type 2 diabetes is not explained by anthropometric, lifestyle or genetic risk factors: the EPIC-InterAct study. Diabetologia. 2013;56: 60–69.

22. Raghavan S, Porneala B, McKeown N, Fox CS, Dupuis J, Meigs JB. Metabolic factors and genetic risk mediate familial type 2 diabetes risk in the Framingham Heart Study. Diabetologia. 2015;58: 988–996.

23. Kannel WB, McGee DL. Diabetes and cardiovascular disease. The Framingham study. JAMA. 1979;241: 2035–2038.

24. Rawshani A, Rawshani A, Franzén S, Sattar N, Eliasson B, Svensson A-M, et al. Risk Factors, Mortality, and Cardiovascular Outcomes in Patients with Type 2

Diabetes. N Engl J Med. 2018;379: 633–644.

25. Stern MP. Diabetes and Cardiovascular Disease: The “Common Soil” Hypothesis. Diabetes. 1995;44: 369–374.

26. Khoury MJ, Mensah GA. Genomics and the prevention and control of common chronic diseases: emerging priorities for public health action. Preventing chronic disease. 2005. p. A05.

27. Goldfine AB, Beckman JA, Betensky RA, Devlin H, Hurley S, Varo N, et al. Family history of diabetes is a major determinant of endothelial function. J Am Coll Cardiol. 2006;47: 2456–2461.

28. Yeung EH, Pankow JS, Astor BC, Powe NR, Saudek CD, Kao WHL. Increased risk of type 2 diabetes from a family history of coronary heart disease and type 2 diabetes. Diabetes Care. 2007;30: 154–156.

29. Lloyd-Jones DM, Nam B-H, D’Agostino RB, Sr, Levy D, Murabito JM, et al. Parental Cardiovascular Disease as a Risk Factor for Cardiovascular Disease in Middle-aged Adults. JAMA. 2004. p. 2204. doi:10.1001/jama.291.18.2204

30. Moonesinghe R, Yang Q, Zhang Z, Khoury MJ. Prevalence and Cardiovascular Health Impact of Family History of Premature Heart Disease in the United States: Analysis of the National Health and Nutrition Examination Survey, 2007--2014. J Am Heart Assoc. 2019;8: e012364.

31. Lloyd-Jones DM, Hong Y, Labarthe D, Mozaffarian D, Appel LJ, Van Horn L, et al. Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association’s strategic Impact Goal through 2020 and beyond. Circulation. 2010;121: 586–613.

32. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention. 2019.

33. American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2019. Diabetes Care. 2019;42: S13–S28.

34. National Health and Nutrition Examination Survey: 2007–2008 Data Documentation, Codebook, and Frequencies: Plasma Fasting Glucose and Insulin (GLU_D). Jan 2010 [cited 14 Aug 2019]. Available: https://wwwn.cdc.gov/Nchs/Nhanes/2007-2008/GLU_E.htm#Description_of_Laboratory_Methodology

35. DEMO_I. [cited 29 Mar 2020]. Available: https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DEMO_I.htm

36. How much physical activity do adults need? | Physical Activity | CDC. 9 Jan 2020 [cited 29 Mar 2020]. Available: https://www.cdc.gov/physicalactivity/basics/adults/index.htm

37. Adults EP on DEAT of HBC in, Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). JAMA: The Journal of the American Medical Association. 2001. pp. 2486–2497. doi:10.1001/jama.285.19.2486

38. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Himmelfarb CD, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. Journal of the American College of Cardiology. 2018. pp. e127–e248. doi:10.1016/j.jacc.2017.11.006

39. Wilson PWF, Pencina M, Jacques P, Selhub J, D'Agostino R Sr, O'Donnell CJ. C-reactive protein and reclassification of cardiovascular risk in the Framingham Heart Study. Circ Cardiovasc Qual Outcomes. 2008;1: 92–97.

40. King DE, Mainous AG, Buchanan TA, Pearson WS. C-Reactive Protein and Glycemic Control in Adults With Diabetes. Diabetes Care. 2003. pp. 1535–1539. doi:10.2337/diacare.26.5.1535

41. C-Reactive Protein, Fibrinogen, and Cardiovascular Disease Prediction. N Engl J Med. 2012;367: 1310–1320.

42. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. 2014.

43. NHANES Response Rates and Population Totals. [cited 29 Mar 2020]. Available: https://wwwn.cdc.gov/nchs/nhanes/ResponseRates.aspx

44. Family Health History and Diabetes | CDC. 1 Jul 2019 [cited 29 Mar 2020]. Available: https://www.cdc.gov/genomics/famhistory/famhist_diabetes.htm

45. Heart Disease. [cited 29 Mar 2020]. Available: https://www.cdc.gov/heartdisease/family_history.htm

46. Wilson PWF, D'Agostino RB, Parise H, Sullivan L, Meigs JB. Metabolic Syndrome as a Precursor of Cardiovascular Disease and Type 2 Diabetes Mellitus. Circulation. 2005. pp. 3066–3072. doi:10.1161/circulationaha.105.539528

47. Holmes MV, Pulit SL, Lindgren CM. Genetic and epigenetic studies of adiposity and cardiometabolic disease. Genome Med. 2017;9: 82.

48. Bao X, Borné Y, Johnson L, Muhammad IF, Persson M, Niu K, et al. Comparing the

inflammatory profiles for incidence of diabetes mellitus and cardiovascular diseases: a prospective study exploring the “common soil”hypothesis. Cardiovasc Diabetol. 2018;17: 87.

49. Shore AC, Colhoun HM, Natali A, Palombo C, Östling G, Aizawa K, et al. Measures of atherosclerotic burden are associated with clinically manifest cardiovascular disease in type 2 diabetes: a European cross-sectional study. J Intern Med. 2015;278: 291–302.

50. Einarson TR, Acs A, Ludwig C, Panton UH. Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017. Cardiovasc Diabetol. 2018;17: 1–19.

51. Haffner SM, Stern MP, Hazuda HP, Mitchell BD, Patterson JK. Cardiovascular risk factors in confirmed prediabetic individuals. Does the clock for coronary heart disease start ticking before the onset of clinical diabetes? JAMA. 1990;263: 2893–2898.

52. Rivera NV, Carreras-Torres R, Roncarati R, Viviani-Anselmi C, De Micco F, Mezzelani A, et al. Assessment of the 9p21.3 locus in severity of coronary artery disease in the presence and absence of type 2 diabetes. BMC Med Genet. 2013;14: 11.

53. Zhao W, Rasheed A, Tikkanen E, Lee J-J, Butterworth AS, Howson JMM, et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. Nat Genet. 2017;49: 1450–1457.

54. Centers for Disease Control and Prevention. Diabetes Report Card 2012. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2012.

55. Moonesinghe R, Beckles GLA, Liu T, Khoury MJ. The contribution of family history to the burden of diagnosed diabetes, undiagnosed diabetes, and prediabetes in the United States: analysis of the National Health and Nutrition Examination Survey, 2009--2014. Genet Med. 2018. Available: https://www.nature.com/articles/gim2017238.pdf?origin=ppub

56. Office of Public Health Genomics, Center for Surveillance, Epidemiology and Laboratory Services (CSELS). My Family Health Portrait: A Tool from the Surgeon General. 18 Sep 2018 [cited 8 Sep 2019]. Available: https://phgkb.cdc.gov/FHH/html/index.html

57. Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on family history of Alzheimer’s disease. Transl Psychiatry. 2018;8: 99.

58. Cui P, Ma X, Li H, Lang W, Hao J. Shared Biological Pathways Between Alzheimer’s Disease and Ischemic Stroke. Front Neurosci. 2018;12: 605.

59. Wang K, Gaitsch H, Poon H, Cox NJ, Rzhetsky A. Classification of common human diseases derived from shared genetic and environmental determinants. Nat Genet. 2017;49: 1319–1325.

60. Biobank UK. Protocol for a large-scale prospective epidemiological resource. Protocol No: UKBB-PROT-09-06 (Main Phase). 2007.

61. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12: e1001779.

62. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143: 29–36.

63. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12: 77.

64. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47: 291–295.

65. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50: 1219–1224.

66. Lamy P-J, Trétarre B, Rebillard X, Sanchez M, Cénée S, Ménégaux F. Family history of breast cancer increases the risk of prostate cancer: results from the EPICAP study. Oncotarget. 2018;9: 23661–23669.

67. Chen Y-C, Page JH, Chen R, Giovannucci E. Family history of prostate and breast cancer and the risk of prostate cancer in the PSA era. Prostate. 2008;68: 1582–1591.

68. Rasooly D, Patel CJ. Conducting a Reproducible Mendelian Randomization Analysis Using the R Analytic Statistical Environment. Curr Protoc Hum Genet. 2019;101: e82.

69. Smith GD, Ebrahim S. Data dredging, bias, or confounding. BMJ. 2002;325: 1437–1438.

70. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol. 2004;33: 30–42.

71. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG, EPIC- InterAct Consortium. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. Eur J Epidemiol. 2015;30: 543–552.

72. Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. Stat Sci. 1999;14: 29–46.

73. Davey Smith G, Ebrahim S. Mendelian randomization: genetic variants as instruments for strengthening causal inference in observational studies. Bio-social surveys: current insight and future promise The National Academies Press, National Research Council, Washington, DC. 2007.

74. Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol. 2018;47: 358.

75. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23: R89–98.

76. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008;27: 1133–1163.

77. Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. Int J Epidemiol. 2016;45: 908–915.

78. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr. 2016;103: 965–978.

79. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7. doi:10.7554/eLife.34408

80. Hemani G, Haycock P, Zheng J, Gaunt T, Elsworth B. TwoSampleMR: Two Sample MR functions and interface to MR Base database.

81. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian randomization. Genet Epidemiol. 2016;40: 597–608.

82. Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. Int J Epidemiol. 2016;45: 1717–1726.

83. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol. 2017;32: 377–389.

84. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol. 2016;40: 304–314.

85. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J

Epidemiol. 2015;44: 512–525.

86. Burdett T, Hastings E, Welter D, SPOT, EMBL-EBI, NHGRI. GWAS Catalog. [cited 13 Dec 2017]. Available: https://www.ebi.ac.uk/gwas/downloads/summary-statistics

87. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in\ 700,000 individuals of European ancestry. bioRxiv. 2018; 274654.

88. Teumer A. Common Methods for Performing Mendelian Randomization. Front Cardiovasc Med. 2018;5: 51.

89. Palmer TM, Sterne JAC, Harbord RM, Lawlor DA, Sheehan NA, Meng S, et al. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. Am J Epidemiol. 2011;173: 1392–1403.

90. Burgess S, Thompson SG, CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. Int J Epidemiol. 2011;40: 755–764.

91. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res. 2017;26: 2333–2355.

92. Wickham H, Hester J, Chang W. devtools: Tools to Make Developing R Packages Easier. 2018. Available: https://CRAN.R-project.org/package=devtools

93. Hemani' 'gibran. MRInstruments: Data sources for genetic instruments to be used in MR. Available: https://github.com/MRCIEU/MRInstruments

94. Wickham H. tidyverse: Easily Install and Load the "Tidyverse." 2017. Available: https://CRAN.R-project.org/package=tidyverse

95. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010;42: 937–948.

96. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium, Mahajan A, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet. 2014;46: 234–244.

97. Thomas DC, Conti DV. Commentary: the concept of "Mendelian Randomization." International journal of epidemiology. 2004. pp. 21–25.

98. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet. 1986;1: 507–508.

99. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC), Wensley F, Gao P, Burgess S, Kaptoge S, Di Angelantonio E, et al. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ. 2011;342: d548.

100. Holmes MV, Lange LA, Palmer T, Lanktree MB, North KE, Almoguera B, et al. Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis. Am J Hum Genet. 2014;94: 198–208.

101. Bennett DA, Holmes MV. Mendelian randomisation in cardiovascular research: an introduction for clinicians. Heart. 2017;103: 1400–1407.

102. Tillmann T, Vaucher J, Okbay A, Pikhart H, Peasey A, Kubinova R, et al. Education and coronary heart disease: mendelian randomisation study. BMJ. 2017;358: j3542.

103. Holmes MV, Dale CE, Zuccolo L, Silverwood RJ, Guo Y, Ye Z, et al. Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. BMJ. 2014;349: g4164.

104. Lawlor D, Richmond R, Warrington N, McMahon G, Davey Smith G, Bowden J, et al. Using Mendelian randomization to determine causal effects of maternal pregnancy (intrauterine) exposures on offspring outcomes: Sources of bias and methods for assessing them. Wellcome Open Res. 2017;2: 11.

105. Lee HA, Park EA, Cho SJ, Kim HS, Kim YJ, Lee H, et al. Mendelian randomization analysis of the effect of maternal homocysteine during pregnancy, as represented by maternal MTHFR C677T genotype, on birth weight. J Epidemiol. 2013;23: 371–375.

106. Swerdlow DI, Preiss D, Kuchenbaecker KB, Holmes MV, Engmann JEL, Shah T, et al. HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. Lancet. 2015;385: 351–361.

107. Åsvold BO, Bjørngaard JH, Carslake D, Gabrielsen ME, Skorpen F, Smith GD, et al. Causal associations of tobacco smoking with cardiovascular risk factors: a Mendelian randomization analysis of the HUNT Study in Norway. Int J Epidemiol. 2014;43: 1458–1470.

108. VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. Methodological challenges in mendelian randomization. Epidemiology. 2014;25: 427–435.

109. Burgess S. Statistical issues in Mendelian randomization: use of genetic instrumental variables for assessing causal associations. University of Cambridge. 2012. Available: https://www.repository.cam.ac.uk/handle/1810/242184

110. Zheng J, Baird D, Borges M-C, Bowden J, Hemani G, Haycock P, et al. Recent Developments in Mendelian Randomization Studies. Curr Epidemiol Rep. 2017;4: 330–345.

111. Verbanck M, Chen C-Y, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet. 2018;50: 693–698.

112. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. Int J Epidemiol. 2016;45: 1961–1974.

113. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Stat Med. 2017;36: 1783–1802.

114. Bowden J, Spiller W, Del Greco M F, Sheehan N, Thompson J, Minelli C, et al. Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. Int J Epidemiol. 2018. doi:10.1093/ije/dyy101

115. van Kippersluis H, Rietveld CA. Pleiotropy-robust Mendelian randomization. Int J Epidemiol. 2017. doi:10.1093/ije/dyx002

116. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. Epidemiology. 2017;28: 30–42.

117. Kahn SE, Cooper ME, Del Prato S. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. Lancet. 2014;383: 1068–1083.

118. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. Diabetologia. 1999;42: 139–145.

119. Chen S, de Craen AJM, Raz Y, Derhovanessian E, Vossen ACTM, Westendorp RGJ, et al. Cytomegalovirus seropositivity is associated with glucose regulation in the oldest old. Results from the Leiden 85-plus Study. Immun Ageing. 2012;9: 18.

120. Sun Y, Pei W, Wu Y, Yang Y. An Association of Herpes Simplex Virus Type 1 Infection With Type 2 Diabetes. Diabetes Care. 2005;28: 435–436.

121. Ioannidis JPA, Loy EY, Poulton R, Chia KS. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. Sci Transl Med. 2009;1: 7ps8.

122. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005;2: e124.

123. Ioannidis JPA, Tarone R, McLaughlin JK. The False-positive to False-negative Ratio in Epidemiologic Studies. Epidemiology. 2011;22: 450–456.

124. de Macedo GMC, Nunes S, Barreto T. Skin disorders in diabetes mellitus: an epidemiology and physiopathology review. Diabetol Metab Syndr. 2016;8: 63.

125. Kornum JB, Thomsen RW, Riis A, Lervang H-H, Schønheyder HC, Sørensen HT. Type 2 diabetes and pneumonia outcomes: a population-based cohort study. Diabetes Care. 2007;30: 2251–2257.

126. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS One. 2010;5: e10746.

127. Ferrannini E, Nannipieri M, Williams K, Gonzales C, Haffner SM, Stern MP. Mode of onset of type 2 diabetes from normal or impaired glucose tolerance. Diabetes. 2004;53: 160–165.

128. Fda US. Drug Administration National Drug Code Directory. Internet address: http://www fda gov/cder/ndc/ (Accessed 2003). 2013.

129. Ho D, Imai K, King G, Stuart E. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, Articles. 2011;42: 1–28.

130. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. 1995;57: 289–300.

131. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47: D1005–D1012.

132. MR-Base. [cited 15 Nov 2017]. Available: http://www.mrbase.org/

133. Soranzo N, Sanna S, Wheeler E, Gieger C, Radke D, Dupuis J, et al. Common Variants at 10 Genomic Loci Influence Hemoglobin A1C Levels via Glycemic and Nonglycemic Pathways. Diabetes. 2010. pp. 3229–3239. doi:10.2337/db10-0502

134. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010;42: 105–116.

135. Mayerle J, den Hoed CM, Schurmann C, Stolk L, Homuth G, Peters MJ, et al. Identification of genetic loci associated with Helicobacter pylori serologic status.

JAMA. 2013;309: 1912–1920.

136. Tian C, Hromatka BS, Kiefer AK, Eriksson N, Noble SM, Tung JY, et al. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat Commun. 2017;8: 599.

137. Johnson EO, Hancock DB, Gaddis NC, Levy JL, Page G, Novak SP, et al. Novel genetic locus implicated for HIV-1 acquisition with putative regulatory links to HIV replication and infectivity: a genome-wide association study. PLoS One. 2015;10: e0118149.

138. Rauch A, Kutalik Z, Descombes P, Cai T, Di Iulio J, Mueller T, et al. Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. Gastroenterology. 2010;138: 1338–45, 1345.e1–7.

139. Rubicz R, Yolken R, Drigalenko E, Carless MA, Dyer TD, Kent J Jr, et al. Genome-wide genetic investigation of serological measures of common infections. Eur J Hum Genet. 2015;23: 1544–1548.

140. Rodrigues CF, Rodrigues ME, Henriques M. Candida sp. Infections in Patients with Diabetes Mellitus. J Clin Med Res. 2019;8. doi:10.3390/jcm8010076

141. Wit SD, De Wit S, Sabin CA, Weber R, Worm SW, Reiss P, et al. Incidence and Risk Factors for New-Onset Diabetes in HIV-Infected Patients: The Data Collection on Adverse Events of Anti-HIV Drugs (D:A:D) Study. Diabetes Care. 2008. pp. 1224–1229. doi:10.2337/dc07-2013

142. He C, Yang Z, Lu N-H. Helicobacter pylori infection and diabetes: is it a myth or fact? World J Gastroenterol. 2014;20: 4607–4617.

143. Chen Y, Blaser MJ. Association between gastric Helicobacter pylori colonization and glycated hemoglobin levels. J Infect Dis. 2012;205: 1195–1202.

144. Jeon CY, Haan MN, Cheng C, Clayton ER, Mayeda ER, Miller JW, et al. Helicobacter pylori infection is associated with an increased rate of diabetes. Diabetes Care. 2012;35: 520–525.