THE UNIVERSITY OF CHICAGO

# COMPUTATIONAL ANALYSIS OF SHARED ETIOLOGICAL FACTORS OF MULTIPLE COMPLEX DISEASES: MINING TEXT AND MEDICAL RECORDS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF GENETICS, GENOMICS AND SYSTEMS BIOLOGY

BY
XIKUAN WANG

CHICAGO, ILLINOIS
AUGUST 2017

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

First, I would like to thank my advisor, Andrey Rzhetsky, for being the luminary on my journey of scientific pursuit. Not only did he introduce me to many fascinating research challenges, he has also given me the academic and emotional support to accomplish them as well as the freedom to learn, explore and grow. Andrey's patient yet persistent stance towards scientific quest, exquisite taste in choosing research topics and ingenious approach to reach elegant solutions are what I hope to emulate.

I have had the honor and pleasure to do much of my research in collaboration with others in the lab and beyond. It is my pleasure to work with present and past members of the Rzhetsky lab, David Blair, Ryan Mork, Shi Yu, Genjie Jia, Rachel Melamed and Ishanu Chattopadhyay. I would also like to thank our collaborators that contributed to the work presented below, including Steven Bagley, Christopher Lyttle, Edwin Cook, Russ Altman, Robert Gibbons, Kelsey Wheeler, Fatima Clouser, Ashtyn Dixon, Kamal El Bissati, Ying Zhou, Rima McLeod, Paulette Krishack, Julian Solway, Anne Sperling, Philip Verhoef, Hallie Gaitsch,Nancy Cox, Hoifung Poon, Robert Stevens, Larisa Soldatova, Ross King, Sophia Ananiadou, Maolin Li, Fenia Christopoulou, Jose Luis Ambite, Sahil Garg, Ulf Hermjakob, Daniel Marcu, Emily Sheng, Aram Galstyan, and James Evans. Also, my committee members, Barbara Stranger, Dan Nicolae, Robert Grossman, and Nancy Cox, have given me their valuable input and advice over the years.

I would also like to thank my friends for being on this journey with me. My "classmates", Yi Zeng, Quan Gao, Pengyao Jiang, Ziyue Gao and Shan Yu have been my surrogate family in Chicago. Of course, I could not do any of this without the support from my parents, Binsheng and Honghong Wang. I internalized the value of lifelong learning by observing them. I'm also grateful to my fiancée, Xiaochen Shi, for her patience, tolerance and understanding. Whenever I felt overwhelmed working on the projects, she was always there for me, bringing joy and stability in my life.

# ABSTRACT

Personalized and precision medicine is a global challenge that requires clear understanding of etiology of common human diseases and complex traits. The rise of data-intensive research and recent advances in genotyping, sequencing and other systems approaches have created an unprecedented opportunity to develop accurate medical diagnosis, efficient therapeutic interventions and cost-effective preventive care. However, modernization of disease classification system that properly integrates our understanding of causal molecular and environmental factors is still in an early stage. In Chapter 1, I will briefly outline recent progresses and challenges associated with our understanding of etiology of common human diseases and complex traits. In Chapter 2 of this dissertation, I will describe a series of studies that utilized the large diagnostic data available through health insurance claims of 150 million patients to discover genetic and especially, environmental factors, that have significant effects on multiple diseases. In Chapter 3, I will illustrate an accurate and reliable classification of complex diseases based on common genetic or environmental factors using in-depth data of about half a million patients. The same analysis can also be used to quantify the genetic and environment effects of hundreds of diseases. In Chapter 4, I will introduces a carefully designed formal ontology and a corpus consists of two significant annotations of biomedical text aimed to facilitate rich digital phenotyping. Finally, in Chapter 5, I will summarize these results and describe how the integration of diverse data in medicine could lead to accurate and precise disease classification and digital phenotyping systems that will transform medical research and patient care.

# CHAPTER 1

# INTRODUCTION

Clear understanding of etiology of common human diseases and complex traits, while currently lacking, in the near future promises to guide us to development of efficient therapeutic interventions and cost-effective disease-prevention measures. Ideally, scientists and medical practitioners would be guided by a comprehensive map of how individual human genotypes interact with various environmental stimuli and result in the phenotypic variations that we observe as individual traits in health and disease.

Advances in genotyping and sequencing technologies brought forth the key findings of the genetic architecture of complex traits, suggesting disease-predisposing genetic variants, both common and rare. Such variants include major, large-scale chromosomal rearrangements (inversions, translocations, chromosome fusions, duplications, deletions), and relatively minor, small-scale events such as single nucleotide polymorphisms (SNPs) and copy number variants (CNVs).

A plethora of studies focused on determining allele frequencies in populations, their effect sizes with respect to human traits. A limited success was achieved in modeling genetic penetrance, most notably in rare diseases. In the case of complex phenotypes the current modeling, purely out of mathematical feasibility, often restricts polymorphisms to being additive, dominant, or recessive. Much effort is applied to delineating patterns of gene–gene and gene–environment interaction effects in human diseases, and discovering genetic pleiotropy or epistasis instances. Just as genes and proteins do not work in isolation from each other [240, 273, 104], human phenotypes can be causaly connected to each other or can be caused by common genetic or environmental factors [99, 119, 130], which manifests in disease comorbidity. *Comorbidity* is clinically defined as set of diseases that tend to co-occur with a primary disease of interest in the same patient, in the excess to expected by chance based on marginal frequencies of diseases in the target population. More broadly it is used to represent the coexistence of two or more diseases in the same individual, sometimes called

*multimorbidity*.

In this work we use the broader definition of comorbidity without prioritizing one disease over another. Since the term *comorbidity* was coined in 1970s [78], comorbidity patterns have been found to be prevalent, using both network approaches and diseases' temporal trajectories [119, 141]. In addition, many studies have attempted to understand the etiological relationship between comorbid diseases, illustrating how susceptible genetic variants combined with environmental perturbation lead to multiple diseases.

Genome-wide association studies(GWAS) have been generally successful in finding genetic variants associated with complex traits and diseases, shedding light on molecular and physiological mechanisms [256, 85, 269, 83]. So far, most GWASs focused on single trait or disease and, occasionally, a group of closely related phenotypes, in order to organize large-scale collaborations and consorted efforts on collecting samples sufficient for detecting significant signals. Recently, there has been a multitude of studies that focused on leveraging the advances in genotyping, sequencing and clinical phenotyping to study etiological relationships of multiple diseases. For example, Consortia such as eMerge, i2b2, deCODE Genetics, UK Biobank and BioBank Japan started to build the foundation for successful multiple-disease studies such as, PheWAS [62], Phenome-wide heritability [92], and pleiotropy studies [109, 77], mapping vast number of genetic variants to a web of comorbid diseases. At the same time, numerous methods have been developed to repurpose publicly available summary statistics to estimate heritability explained by SNPs [92, 36, 325] and genetic correlations across diseases and traits [52, 35, 175, 7]. This group of studies probe into the promising future where the connections between genotypes and phenotypes are understood as a network-to-network mapping.

Unfortunately, three serious roadblocks still stand between a clear understanding of the complex relationship between genotypes, environments and phenotypes. First, although a wealth of genotypic information are publicly available and advance tools are being developed to leverage it, environmental and epidemiological studies are lagging behind in understanding

environmental patterns of multiple correlated diseases. Second, confined by the phenotypes they focused on, progresses made by genetic and genomic research, especially by those across multiple diseases, have limited impact on current classification of diseases, which is heavily based on human anatomy, clinical manifestations, and heuristics. Third, because phenotype data are stored in multiple formats catering to various purposes, fragmented phenotypic information is becoming a bottleneck in connecting genomic, transcriptomic and proteomic networks to the phenotypic network. Below, I briefly review each of these challenges in greater detail along with the progresses that have been made with genomic analyses, epidemiological studies and clinical phenotypic data.

### *Discovering Etiological Factors Driving Multiple Diseases*

Epidemiology studies such as Ni-Hon-San study [147] and the Irish-Brothers Study [288] demonstrated significant and sometimes unexpected effects of environments on diseases. Since then, studies found significant etiological effects of BMI, smoking, diet, exercise and occupational exposures. Epidemiology studies have successfully linked causal environmental factors to specific diseases such as smoking to lung cancer [65], H. pylori infection to peptic ulcer and stomach cancer [277], HPV infection to cervical cancer [199], hepatitis C infection to liver cancer [258], and alcohol to liver cirrhosis [208].

More importantly, they showed that important adult risk factors for chronic diseases can be traced back to childhood or early adult environments. Adult allergies, for example, are significantly influenced by early life exposure or a lack of exposure. Early life exposure to viral (i.e. HPV), bacterial (i.e. Helicobacter pylori) and parasitic (i.e. toxoplasmosis) infections can have significant effects on late-onset or chronic disease [277, 199, 27]. Furthermore, permanent developmental changes may be caused by prenatal environmental exposure during critical and sensitive periods of fetal development. Later-life events may have limited effects comparing to these sensitive periods [161, 188]. Other traits may be more resilient to early environmental assaults [14, 72].

Similar to expanding research on the pleiotropic effects of genetic variants, increasing number of studies has focused on finding common environmental factors causal to multiple diseases. For example, environmental assaults such as infections can have a significant effect across multiple biological systems and result in various phenotypic manifestation, providing key insights on the development of autoimmune diseases and neuropsychiatric diseases[40, 71]. Common environmental effects are also prominent in allergies, such as allergic rhinitis, asthma, and chronic sinusitis[43, 176, 53]. In addition, chemicals and drug intakes, especially those disrupt microbiomes, frequently encompass correlated health outcomes, while socioeconomic and cultural environments could also affect the clustering of multiple diseases. Life style factors such as diet, smoking and alcohol have been identified as risk factors linked to common cancers and cardiovascular diseases [152]. Sunlight exposure affect human physiology in multiple ways mostly via vitamin D levels. In addition to its significance in maintaining calcium homeostasis and bone related health, Vitamin D was shown to have immuno-modulatory properties and implicated in the pathogenesis of autoimmune disorders. Some reports imply that in multiple disorders such as Parkinson's disease, multiple sclerosis, and diabetes type 1, vitamin D may even be preventive [187, 9, 68].

In Chapter 2 of this dissertation, I will describe a series of studies that utilize the large diagnostic data available through health insurance claims to discover genetic and especially, environmental factors, that have significant effects on multiple diseases. Specifically, in section 2.1, I describe a study that used an indirect approach to represent environmental factors and uncovered their effects on the geographic patterns of autism and intellectual disability incidences. In section 2.2, I describe a study focusing on the effect of a specific environmental assault, toxoplasmosis, on multiple biological systems, causing a array of comorbid diseases. In section 2.3, I introduce a study focused on comorbid patterns of neuropsychiatric patients and possible underlying factors driving these patterns. In section 2.4, I describe a study that used a mouse model analysis to test a hypothesis generated from comorbid patterns of sepsis patients to reveal a novel immunologic mechanism.

## *Quantifying Genetic and Environmental Contributions across Multiple Diseases*

Since its proposal by Fisher almost a century ago [80], the infinitesimal model has formed the conceptual foundation in understanding genetic architecture of complex traits. Similarly, disease liability [73] are most likely affected by a large number of loci that each has small effect.In order to understand complex trait genetic architecture, GWAS have successfully identified genetic variants in millions of individuals across hundreds of complex traits while partitioning the genetic and environmental effects on complex diseases and demonstrating that GWAS variants explain significant portion of narrow-sense heritability for single diseases [220].

To study pleiotropy and multi-functionality, multiple research efforts have been set up to detect genetic variants and environmental factors that are otherwise elusive using single-disease studies [130]. Shared etiology among multiple diseases, either via common genetic variants [220] or common environmental factors has been pursued in correlation studies. Genetic correlation, that is the proportion of variance that two traits share due to genetic causes, indicates the strength of the genetic factor common between two traits and its direction. Genetic correlations can arise through multiple mechanisms [130], most of which indicate same biological process underlying multiple diseases. Similarly, environmental correlation, the counterpart of genetic correlation, could also shed light on common biological or molecular processes of multiple disease development [180]. Understanding of the environmental and genetic correlations of multiple diseases often requires collaborations between studies and consortiated efforts [52, 276, 35, 180]. For example, UK biobank recruited 500,000 participants with age ranging from 40 to 69, for a prospective study with 20 years of follow-up. Participants have undergone detailed phenotyping, and multiple factor of life styles and medical outcomes are recorded. On the other hand, the groundbreaking whole-country analysis of family-based genetics of schizophrenia and bipolar disorder in Sweden [169] demonstrated the validity of using a health-registry-based retrospective study to un-

cover genetic and environmental effects on multiple diseases. Obviously, this approach has its limitation, mainly due to the difficulties in acquiring phenotypic data from patients. However, it illustrates that family based analyses incorporated with large phenotypic data can provide significant power in differentiating genetic and environmental effects on the complex relationship between multiple diseases.

Studies like the meta-analysis of the heritability of all human traits published [226], gave us a glimpse of the future where information across traits are incorporated together to form a much complete map. Unfortunately, similar studies of multiple traits across different systems are much harder to scale than single trait studies. In Chapter 3 of this dissertation, I will describe a study that attempted to quantify the genetic and environment effects on 149 diseases across 20 biological systems. This study utilized inferred family relationships of a large number of patients (over 150 million) to conduct family based analyses on the heritability of, and the genetic and environmental correlations between these diseases.

### *Annotating Textual Corpus for Digital Phenotyping*

Phenotyping studies expand our knowledge on phenotype definitions and disease classification, as a trait's definition is relative to its harboring system. This is even more crucial when we study the relationship between diseases. Clear definition of diseases helps us distinguish 'true' pleiotropic genetic variants or multi-causal environmental factors from 'mediated' relationships of multiple diseases. While digital genotyping is prevalent, as discussed above, digital phenotyping is lagging behind. Phenotyping is currently crude and majority of the phenotyping relies on structured data, such as diagnostic codes, prescription drug codes and surgical procedure codes. There is an untapped well of clinical information in unstructured clinical text. Unstructured clinical texts such as medical narratives are the primary sources of communications between medical professionals. Unstructured clinical text can significantly increase accuracy in phenotyping, especially when linked with structured data such as billing codes and medications [307]. However, [306] also argued that accurate phenotypes

and relationships hidden in unstructured text remain difficult to extract due to the lack of resources describing the semantic relationships between clinical concepts. The richness and unique context of clinical text hinders the performance from general, context-agnostic tools mostly trained on general language corpora, such as Penn TreeBank [186].

A text corpus is usually defined as a collection of documents, annotated according to linguistic rules or domain knowledge of a specialized field. To bridge the gap between patient care and research, various clinical Natural Language Processing (NLP) systems have attempted to extract information from clinical narrative text [304]. For example, MetaMap [10], extensively used and sometimes regarded as the industry standard for mapping medical terms, is a biomedical NLP system built to extract concepts in Unified Medical Language System (UMLS) MetaThesaurus [28] from literature. Still, the performance of Metamap leaves much to be desired. Even with strong software and mapping between medical concepts and literature, existing knowledge bases in the medical domain do not have good coverage on a complex system of phenotypes and biomedical narratives. In addition, one of the major hurdles in leveraging the current progress in molecular biology and genotyping studies for clinical usage is a lack of systematic way to connect molecular pathways and clinical features [205]. Therefore, an annotated biomedical corpus is crucial in the improvement of the current digital phenotyping system, especially on tasks such as named-entity recognition (NER) and subsequently event extraction. Furthermore, an annotated corpus can be used to train algorithms and benchmark different systems of NLP tools.

In chapter 4, I will describe a study aimed to improve digital phenotyping by combining rich annotation of biomedical text with domain knowledge stored in formal ontologies. Specifically, this study introduces a corpus consists of text from literature published in biomedical fields and two significant annotations, term annotation and event annotation, plus the relationships between the annotations. Both annotations are distinct from regular language annotation in that it is biologically and medically meaningful. We focused not only on the linguistic structure but also on the biological and medical meanings. Our annotation

also aims at establishing the ontology unifying phenotypes, genes, proteins, drugs, chemicals, treatments, and procedures that can help us better utilize information from clinical texts.

# CHAPTER 2

# DISCOVERING ETIOLOGICAL FACTORS DRIVING MULTIPLE DISEASES

## 2.1 Introduction

Human diseases, especially neuropsychiatric disorders, are highly comorbid [98, 294, 130]. Leveraging the large scale diagnostic data, we applied the patient matching and geographic methods to identify comorbid patterns for specific patient groups. This disease comorbidity analysis could shed new light onto the etiology of multiple diseases and enable novel genetic and environmental analyses. In the first section of this chapter, I describe a study that used approximate measurements for exposure to environmental toxins and uncovered their effects on autism and intellectual disability incidences. In section 2.2, I describe a study estimating national incidences of a specific environmental assault, toxoplasmosis, and its significant effects on multiple biological systems. In section 2.3, I introduce a study focusing on comorbidity patterns of patients of 32 Neuropsychiatric diseases and infer etiological factors driving these comorbidities. In section 2.4, I describe a study that used a mouse model to validate the inverse comorbid patterns between sepsis and allergic diseases to reveal a novel immunologic mechanism.

## 2.2 Environment and incentives affect the incidence of autism and intellectual disability

### *2.2.1 Introduction*

Autism spectrum disorders (ASD) are a collection of chronic, complex neuropsychiatric diseases with well-characterized comorbidities and increasing apparent prevalence [51]. With few and limited effective treatments and considerable financial burden, its etiology remains a scientific puzzle. Evidence suggests that Autism is highly heritable and clustered within families; consequently, much scientific attention has been dedicated to the discovery of predisposing genetic factors [243, 25, 39, 251, 138, 69]. There is also evidence for environmental influences, such as prenatal exposure to pesticides or valproate, but it is challenging to account systematically for these factors because they are mostly undocumented. In addition, there are numerous factors that could affect or distort the observed variation in temporal and spatial disease prevalence: evolving diagnostic criteria, socioeconomic, legal, and cultural incentives for diagnosis [104], changing environmental exposures, and the accumulation of genetic burden in the growing human population. However, the relative importance of all these putative causal factors and confounders on ASD prevalence, the nature of interactions between contributing factors, and the underlying biological mechanisms, remain unclear.

Along these lines, geospatial clustering of ASD has been observed in California [150, 296], Texas [167, 275], North Carolina [120] and Utah [223]. Clustering could indicate the existence of localized risk factors, such as environmental toxins [20] or maternal education [139]. However, studies to date have focused primarily on within-state patterns and socioeconomic predictors, such as the level of parental education, the controversial financial incentives induced by state policies for special-education services, and broad environmental indicators. Now, the increasing availability of large administrative clinical datasets, with national coverage and fine spatial granularity, along with data on possible causal and confounding factors, provides an opportunity to compare the magnitudes of these and other factors within a

unified mathematical framework.

Here, we report a mixed-effect Poisson regression analysis of the spatial incidence patterns of ASD and, for comparison, intellectual disability (ID). The data was derived from a very large insurance claims database containing nearly 100 million patients in the United States, which was augmented with census data to introduce additional county-level covariates that captured socioeconomic, demographic, and environmental effects. We present strong statistical evidence for environmental and legal factors driving the apparent spatial heterogeneity of both phenotypes, while documented socioeconomic factors and population structure have much weaker effects.

### *2.2.2 Results*

We analyzed the strength of disparate factors on the apparent incidence rates of ASD and ID by computationally interrogating insurance claims for approximately one third of the US population, using a bivariate-response, three-level, mixed-effects Poisson regression model with 50 free parameters, 44 of which correspond to the fixed effects of known factors while the remaining 6 account for the variance and covariance among the random effects (see Methods). The bivariate outcomes modeled by these parameters were the incidence counts for the ASD and ID phenotypes, tabulated separately for males and females in 3,111 counties, nested within 50 states (plus the District of Columbia) and adjusted for population size.

The results are summarized in Figures figs. 2.1 to 2.5 and Table 2.2: We observed clear spatial clusters for both ASD and ID. The raw data analyzed prior to complex modeling (see Figure 2.1) indicated that putative environmental variables were strongly predictive of rates of ASD and ID across the USA. This trend persisted after the analysis was corrected for confounding variables using mixed-effect Poisson regression, see Figures figs. 2.2 to 2.5. We found that ASD in males (normalized by county population) has a county-level mean rate of 0.1% per male of any age. The distribution of rates across counties is skewed: the median is 0.023% while maximum observed value is 5.2%. Similarly, for ID, the average rate over the

whole country is 0.024% per any male, 0.025% per any female. The maximum per county per person rate reached 0.9% and 0.58% for males and females, respectively. Furthermore, for males, the rates of ASD and ID at the county level were weakly but significantly correlated (Pearson product-moment correlation 0.0589, p=0.00101), while for females the correlation was much stronger (0.197, $p < 2.26 \times 10^{-16}$). The estimated Poisson rates of incidence produced by model-based inference are shown in Figure 2.2. When direct comparison was possible, we compared our conclusions with those of prior studies ([150, 296, 167, 275, 120, 223], see Table 2.3) and found that they were consistent.

**Table 2.1: Markov chain Monte Carlo estimates of regression weights, corresponding event rate ratios (exponential of regression weights).** and 95% event estimate credible intervals. The 3-level Poisson mixed-effect model included 12,444 level-1 units (incidence counts for two diseases), 3,111 Level-2 units (counties), and 51 level-3 units (states). The table is designed to mirror Figure 2.1: fixed-effect parameter estimates for autism spectrum disorders (AU) are followed by the corresponding estimates for intellectual disability (ID). For each group of diseases the first parameter listed (AU and ID, respectively) is the intercept. The rest of the fixed-effect parameters are ordered from the strongest negative effect to the strongest positive effect, as in Figure 2.1.

| Parameter | Effect | Estimate | Event Rate | Estimate CI (L U) | | Rate CI (L U) | | p-value |
|---|---|---|---|---|---|---|---|---|
| AU | | 0.367 | 1.443 | -0.973 | 1.755 | 0.378 | 5.782 | 0.56225 |
| AU:Eval1 | - | -4.231 | 0.015 | -8.878 | -0.328 | 0.000 | 0.720 | 0.02475 |
| AU:Gender | - | -0.699 | 0.497 | -0.709 | -0.690 | 0.492 | 0.502 | $< 6 \times 10^{-5}$ |
| AU:Pacific | - | -0.439 | 0.645 | -0.701 | -0.173 | 0.496 | 0.841 | 0.0015 |
| AU:Eval2 | - | -0.213 | 0.808 | -2.922 | 2.471 | 0.054 | 11.830 | 0.883 |
| AU:ASD1 | - | -0.175 | 0.839 | -1.582 | 1.164 | 0.205 | 3.203 | 0.83788 |
| AU:AmInd | - | -0.131 | 0.877 | -0.224 | -0.054 | 0.799 | 0.947 | $< 6 \times 10^{-5}$ |
| AU:B | - | -0.131 | 0.877 | -0.217 | -0.054 | 0.805 | 0.948 | $< 6 \times 10^{-5}$ |
| AU:WHisp | - | -0.128 | 0.880 | -0.213 | -0.053 | 0.808 | 0.949 | $< 6 \times 10^{-5}$ |
| AU:W | - | -0.122 | 0.885 | -0.210 | -0.046 | 0.810 | 0.955 | $< 6 \times 10^{-5}$ |
| AU:CFR1 | - | -0.100 | 0.905 | -1.789 | 1.491 | 0.167 | 4.440 | 0.95962 |
| AU:Asian | - | -0.045 | 0.956 | -0.137 | 0.037 | 0.872 | 1.038 | 0.3455 |

Continued on next page

**Table 2.1 – continued from previous page**

| Parameter | Effect | Estimate | Event Rate | Estimate CI (L U) | | Rate CI (L U) | | p-value |
|---|---|---|---|---|---|---|---|---|
| AU:Poor | - | -0.035 | 0.966 | -0.050 | -0.021 | 0.952 | 0.980 | $< 6 \times 10^{-5}$ |
| AU:Insured | - | -0.012 | 0.988 | -0.021 | -0.003 | 0.979 | 0.997 | 0.00725 |
| AU:Income | + | 0.032 | 1.033 | 0.023 | 0.042 | 1.023 | 1.043 | $< 6 \times 10^{-5}$ |
| AU:Urban | + | 0.036 | 1.037 | 0.034 | 0.038 | 1.035 | 1.039 | $< 6 \times 10^{-5}$ |
| AU:BHisp | + | 0.085 | 1.089 | -0.102 | 0.255 | 0.903 | 1.290 | 0.34987 |
| AU:DSM1 | + | 0.138 | 1.148 | -1.120 | 1.462 | 0.326 | 4.316 | 0.80438 |
| AU:Viral_M | + | 0.180 | 1.197 | 0.122 | 0.239 | 1.130 | 1.269 | $< 6 \times 10^{-5}$ |
| AU:CongMrepM | + | 0.277 | 1.319 | 0.119 | 0.422 | 1.127 | 1.525 | 0.00025 |
| AU:Eval-1 | + | 0.668 | 1.949 | -0.774 | 2.342 | 0.461 | 10.404 | 0.366 |
| AU:ConGenM | + | 1.345 | 3.838 | 0.753 | 1.912 | 2.123 | 6.770 | $< 6 \times 10^{-5}$ |
| | | | | | | | | |
| ID | | -0.227 | 0.797 | -1.394 | 0.980 | 0.248 | 2.665 | 0.66487 |
| ID:Eval1 | - | -4.536 | 0.011 | -8.643 | -1.123 | 0.000 | 0.325 | 0.00637 |
| ID:Pacific | - | -0.374 | 0.688 | -0.612 | -0.121 | 0.543 | 0.886 | 0.00487 |
| ID:AmInd | - | -0.139 | 0.870 | -0.216 | -0.070 | 0.805 | 0.932 | $< 6 \times 10^{-5}$ |
| ID:B | - | -0.130 | 0.878 | -0.206 | -0.066 | 0.814 | 0.936 | $< 6 \times 10^{-5}$ |
| ID:WHisp | - | -0.130 | 0.878 | -0.204 | -0.066 | 0.816 | 0.936 | $< 6 \times 10^{-5}$ |

Continued on next page

**Table 2.1 – continued from previous page**

| Parameter | Effect | Estimate | Event Rate | Estimate CI (L U) | | Rate CI (L U) | | p-value |
|---|---|---|---|---|---|---|---|---|
| ID:W | - | -0.127 | 0.881 | -0.201 | -0.061 | 0.818 | 0.941 | $< 6 \times 10^{-5}$ |
| ID:Gender | - | -0.114 | 0.892 | -0.127 | -0.102 | 0.881 | 0.903 | $< 6 \times 10^{-5}$ |
| ID:ASD1 | - | -0.109 | 0.897 | -1.371 | 1.038 | 0.254 | 2.824 | 0.8965 |
| ID:Asian | - | -0.066 | 0.937 | -0.146 | 0.003 | 0.864 | 1.003 | 0.05613 |
| ID:Eval2 | - | -0.063 | 0.939 | -2.419 | 2.327 | 0.089 | 10.245 | 0.97462 |
| ID:Poor | - | -0.013 | 0.987 | -0.027 | 0.002 | 0.973 | 1.002 | 0.08038 |
| ID:Insured | + | 0.005 | 1.005 | -0.004 | 0.015 | 0.996 | 1.015 | 0.30088 |
| ID:Income | + | 0.027 | 1.028 | 0.018 | 0.037 | 1.018 | 1.037 | $< 6 \times 10^{-5}$ |
| ID:Urban | + | 0.031 | 1.032 | 0.029 | 0.033 | 1.030 | 1.034 | $< 6 \times 10^{-5}$ |
| ID:BHisp | + | 0.033 | 1.033 | -0.129 | 0.195 | 0.879 | 1.215 | 0.66837 |
| ID:CFR1 | + | 0.103 | 1.108 | -1.422 | 1.453 | 0.241 | 4.275 | 0.829 |
| ID:DSM1 | + | 0.149 | 1.161 | -0.972 | 1.286 | 0.378 | 3.617 | 0.76525 |
| ID:Viral_M | + | 0.211 | 1.235 | 0.153 | 0.270 | 1.166 | 1.309 | $< 6 \times 10^{-5}$ |
| ID:CongMrepM | + | 0.362 | 1.437 | 0.210 | 0.516 | 1.234 | 1.675 | $< 6 \times 10^{-5}$ |
| ID:Eval-1 | + | 0.418 | 1.519 | -0.872 | 1.850 | 0.418 | 6.358 | 0.54113 |
| ID:ConGenM | + | 0.660 | 1.935 | 0.014 | 1.255 | 1.014 | 3.509 | 0.03838 |

**Abbreviations** AU – autism spectrum disorders; ID – intellectual disabilities; AmInd – proportion of American Indians; Asian – proportion of Asians; WHisp – White Hispanics; W – White non-Hispanic; BHisp – black Hispnics; B – black non-Hispaic; Pacific – Pacific Islanders; Insured – proportion of insured; Poor – proportion of poor; Urban – proportion of urban; CongMrepM – congenital malformations excluding malformations of genitals in males; ConGenM – congenital malformations of genitals in males; Viral_M – viral infections affecting males; ASD – inclusion of Autism Spectrum Disorders; CFR – Code of Federal Regulations; DSM – requirement of reference to Diagnostic and Statistical Manual of Mental Disorders; Eval – rigor of evaluation of diagnosis veracity. The regulations were encoded in the following way (23). For CFR: Code -1 if criteria included information from the autism section of CFR only, and +1 if the criteria incorporated additional non-CFR information. For DSM: Code -1 if the entire DSM-IV-TR criteria were used, and +1 otherwise. For ASD: -1 if autism spectrum disorders were included in diagnostic criteria, and +1 if they were not included. For Eval: -1 if a diagnosis by a pediatrician or clinician was mandatory, 1 if a diagnosis of autism or autism spectrum disorders by a pediatrician or clinician was mandated, and 2 if no requirements in addition to those mandated by the Individuals with Disabilities Education Act were added.

| Variable | Covariance | Correlation | Cov 95% CI (L U) | |
|---|---|---|---|---|
| AU:AU (State) | 3.126 | 1 | 0.6062 | 7.239 |
| AU:ID (State) | 2.666 | 0.974 | 0.5198 | 6.088 |
| ID:ID (State) | 2.398 | 1 | 0.561 | 5.358 |
| | | | | |
| AU:AU (County) | 0.7699 | 1 | 0.7142 | 0.8285 |
| AU:ID (County) | 0.696 | 0.962 | 0.6462 | 0.7488 |
| ID:ID (County) | 0.6796 | 1 | 0.6261 | 0.7357 |

**Table 2.2: Markov chain Monte Carlo estimates of covariances and correlations of random effects across two phenotypes.**

| | Study(State) | | | | |
|---|---|---|---|---|---|
| | [150](CA) | [296](CA) | [167, 275](TX) | [120](NC) | [223](UT) |
| **Data source** | All live births and diagnostic records for children born in CA between 1992 and 2000. | All live births in CA occurring in 1996–2000 | Administrative educational data for prevalence of autism and other special education categories for the academic years 2000–2001 through 2005 – 2006. | Record-based surveillance for 8 NC counties biennially from the Autism and Developmental Disabilities Monitoring Network. | Record-based surveillance for eight-year-old children born in 1994 and living in Utah in 2002 from the Utah Registry of Autism and Developmental Disabilities Program. |
| Cases + controls | 4,906,926 | 2,453,717 | 4,057,712 | 11,034 | 26,108 |

**Table 2.3: Summary of the prior state-level studies regarding geographic clustering of ASD.**

Continued on next page

| | | | | | |
|---|---|---|---|---|---|
| Cases | 18,731 | 9,900 | 7,022 (ASD+ID) | 532 (ASD), 1,028 (ID) | 99 (ASD-only), 33 (ASD and ID), 113 (ID-only) |
| Modeling formalism | Multilevel logistic regression. | Spatial clustering and bivariate mixed Poisson regression. | Multilevel Poisson regression. | Generalized additive model. | Multiple single-variable logistic regressions. |
| **Factors** | Individual-level Factors: First Born, Premature, Normal Weight, Gender, Low Apgar, | Individual-level Factors: | Environmental Factors (air pollution): | Individual-level Factors: | Individual-level Factors: |

**Table 2.3: Summary of the prior state-level studies regarding geographic clustering of ASD.**

Continued on next page

| | | | | | |
|---|---|---|---|---|---|
| | Max Parental Age, Education, Missing Father. | Mother's Age, Father's Age, the Highest Parental Education Level, Ethnicity. | Mercury, Antimony, Lead, Manganese, Nickel, Zinc, Benzene, Ethylbenzene, Naphthalene , Trichloroethylene, Sulfuric Acid. | Gender, Year of Birth, Plurality of Children, Maternal Age, Ethnicity, Maternal Level of Education, Tobacco Use During Pregnancy, Method of Delivery, Birth Weight, Adequacy of Prenatal Care. | Gender, Mother's Ethnicity, Maternal Age, Paternal Age, Maternal Education, Paternal Education, Adjusted Gross Income,Federal Taxes Paid, Tax Exemptions. |
| Conclusion | ASD are correlated with parental education and income. | ASD is correlated with high parental education. | Mercury-ASD association is uncertain; evidence of Nickel-ASD association. | ASD is associated with maternal age and education. | ASD is associated mothers being of White non-Hispanic ethnicity and older than 34 at the child's birth. |

**Table 2.3: Summary of the prior state-level studies regarding geographic clustering of ASD.**

Continued on next page

| | | | | | |
|---|---|---|---|---|---|
| Consistency with the current ASD geographic clustering study | Not available, as the study did not report disease prevalence data. | Geographic clustering was consistent with our estimates. | The study reported prevalence as quartiles, and data for the majority of counties in Texas was missing, clustering was hard to compare directly. | Highly consistent with our ASD estimates, although only 8 NC counties were analyzed. | No spatial prevalence distribution data was published. |

**Table 2.3: Summary of the prior state-level studies regarding geographic clustering of ASD.**

Continued on next page

Accumulating evidence [292, 30, 87, 89, 249, 123, 122, 234, 142, 165, 139, 118, 170] suggests that the rate of birth malformations, especially of those affecting the reproductive system in newborn boys [30], adjusted for population size and structure, could serve as an indicator of average parental exposure to toxins within a geographic unit. After controlling for ethnicity, gender, and socioeconomic factors, the strongest predictor of ASD was the rate of male congenital malformations of the reproductive system, used as an approximate measurement for exposure to teratogens, based on extensive epidemiological evidence ([292, 30], see Figure 2.3 and Discussion). Every additional percent incidence of male congenital malformations of the reproductive system was predictive of a 283% increase in the rate of the ASD incidence (95% confidence interval, CI: [91%, 576%], p ¡ 6 10-5). Similarly, non-reproductive congenital male malformations accounted for a 31.8% ASD rate increase (CI: [12%, 52%], p ¡ 6 10-5). In contrast, male congenital malformations of the reproductive system were barely significantly predictive for ID (94%, CI: [1%, 250%], p = 0.0383). However, the effect of non-reproductive congenital malformations in males on ID incidence was statistically significant and strong: an increase of 43% (CI: [23%, 67%], p ¡ 6 10-5). Another variable significantly affecting both ASD and ID was population-adjusted incidence of viral infections in males (Table 2.1A, Figure 2.3, and Discussion). Moreover, comorbidity analysis demonstrated that male children with ASD are 5.53 times more likely to have congenital genital malformations than unaffected males (odds ratio 95% CI [5.22, 5.87], p ¡ 2.2 $\times 10^{-16}$, Fisher's exact test).

Male congenital malformations of the reproductive system are subdivided in the ICD9 taxonomy [309], which was used to encode the data in this analysis, into unspecified malformations, hypospadias (abnormally placed external urethral orifice), epispadias (the urethra does not develop into a full tube), micropenis, and undescended testicles. The US average incidence rate for all male congenital malformations of reproductive system was 0.2687% per male of any age group; of these 18% were hypospadias (0.049% rate), 6% congenital chordees (0.0161% rate), 1% micropenis (0.00275% rate), and 0.083% epispadias (0.00227% rate). The rest of the malformations were in the unspecified category; undescended testicles

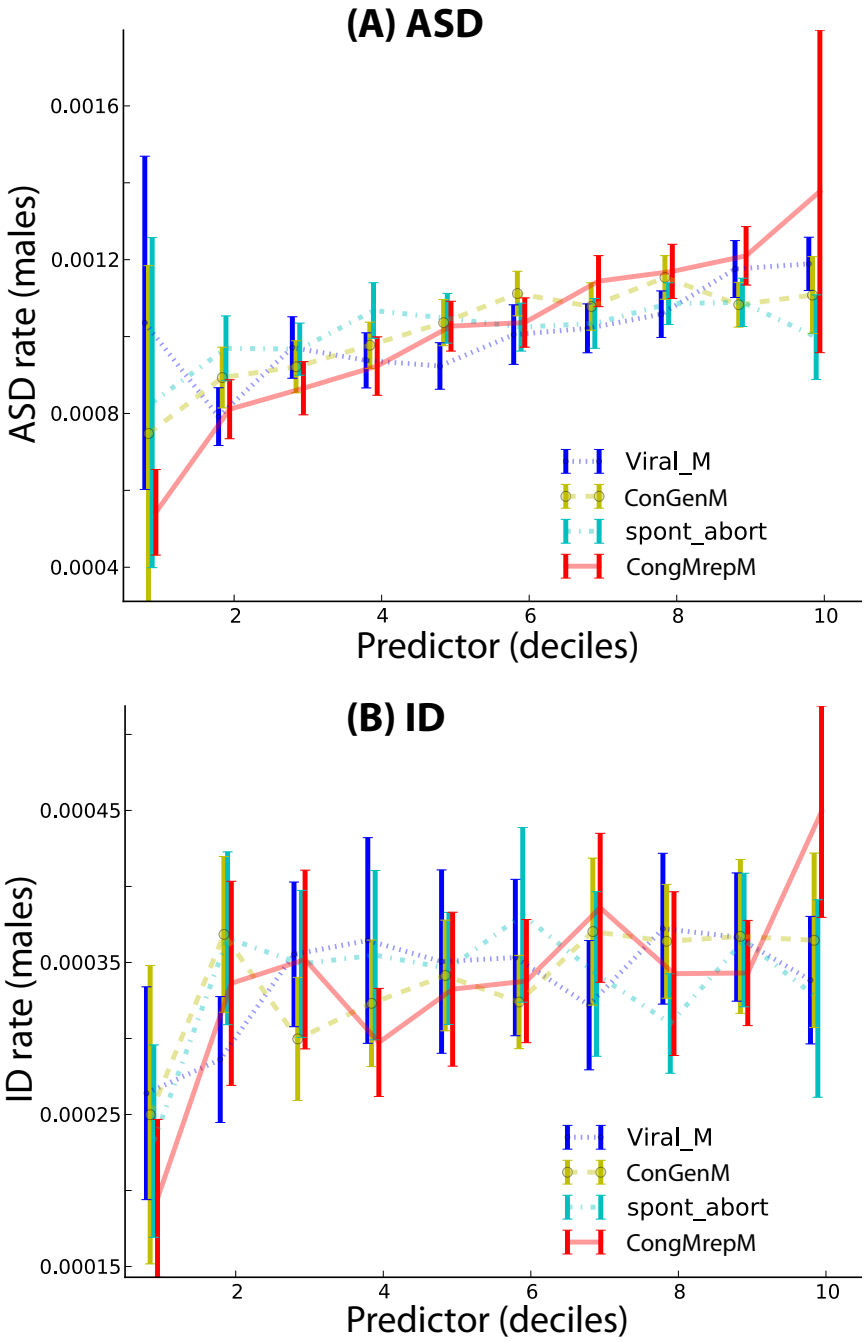


**Figure 2.1: Deciles of putative predictors for ASD (A) and ID (B) per-male rates.** Plots for females look essentially identical, but the absolute rate of incidence is lower (data not shown). The predictor variables shown here are male congenital malformations of reproductive system (ConGenM), viral infections in males of any age (Viral_M), congenital malformations excluding malformations of the genitals in males, CongMrepM, and spontaneous abortion (spont_abort).

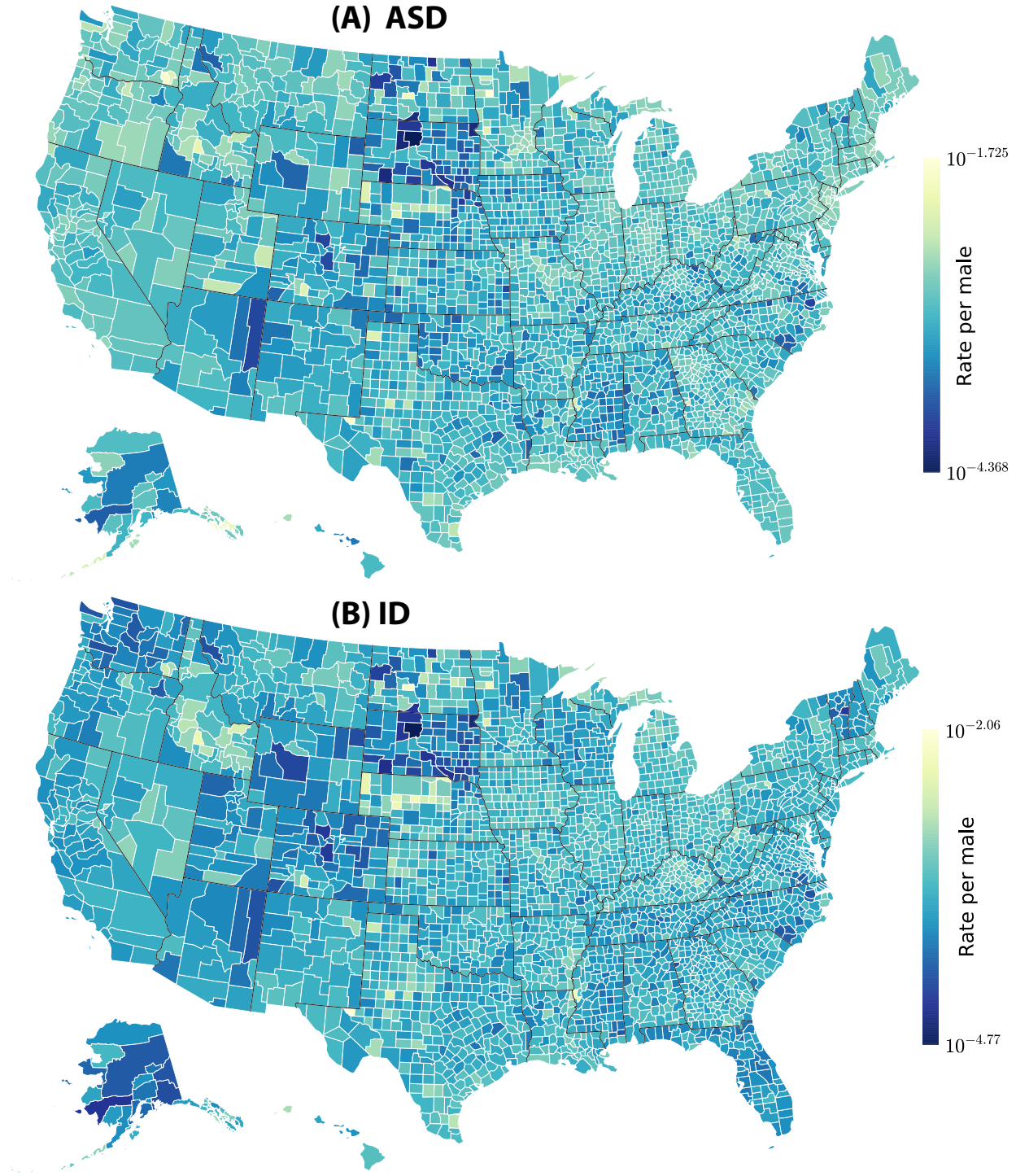


**Figure 2.2: Estimated Poisson rates of incidence of ASD (A) and ID (B) per male individual of any age.**

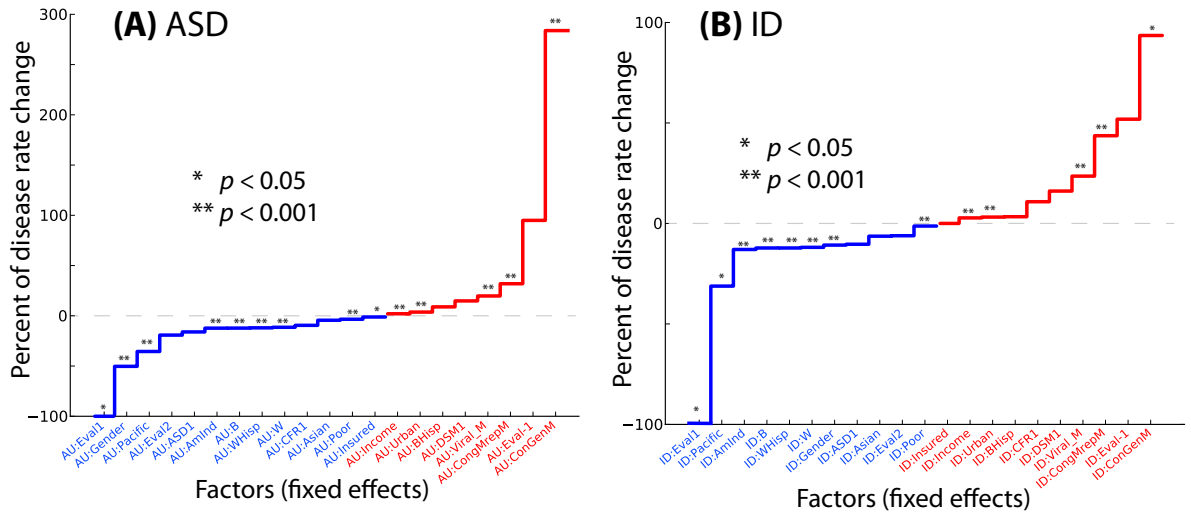


**Figure 2.3: Comparison of fixed effects (geographically varying factors) governing rate variation in ASD (A) and ID (B).** The asterisks indicate the level of significance of individual regression coefficients; see the figure key and Table 1.

were not encoded explicitly. Per-county rates of hypospadias and congenital chordees were significantly correlated with each other (Pearson's r = 0.34, 95% CI [0.31, 0.372], p ¡ 2.2 10-16), as were hypospadias and epispadias (r = 0.066, 95% CI [0.031, 0.101], p = 0.00022). The highest per county rates of malformations were 2.4% (all male genital malformations), 1.1% (hypospadias), 0.91% (chordees), 1.1% (epispadias), and 0.23% (micropenis). There were counties with no reported malformations (zero apparent rate). All discussed groups of malformations were significantly correlated with autism rates. Female birth malformations of reproductive system, variable ConGenF, showed very similar disease-specific predictive behavior to the congenital male malformations of reproductive system, variable ConGenM. The female malformations were predictive of an increase in both ASD and ID, but the magnitude of the statistical effect associated with this factor was much smaller than ConGenM, although highly correlated.

Both ASD and ID showed significant gender-specific incidence effects, with males affected more frequently than females; this was more extreme for ASD (Table 2.2 and Figure 2.3). Using ethnicity variables to account for genetic heterogeneity of the US population, corrected for socioeconomic factors, such as the mean county-specific income, we found the incidence

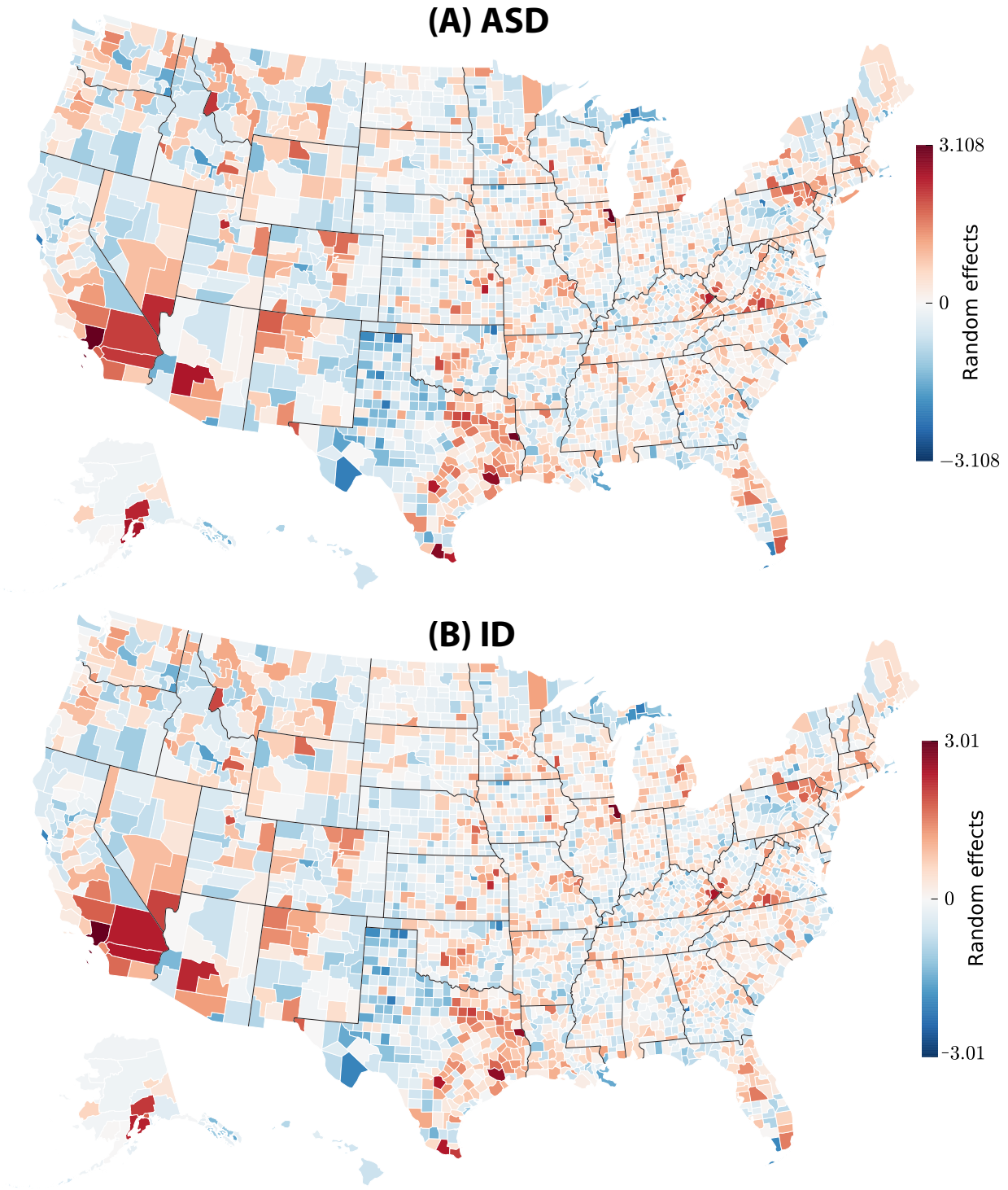


**Figure 2.4: Total state-level random effects of ASD and ID incidence in the USA: (A) ASD and (B) ID.** In the figures we color-coded the Empirical Bayes estimates of the state-level random effects, separately for ASD and ID. County- and state-level random effects model the unknown factors that vary geographically and govern differences in county-specific disease rates after accounting for all fixed effects (see Methods).

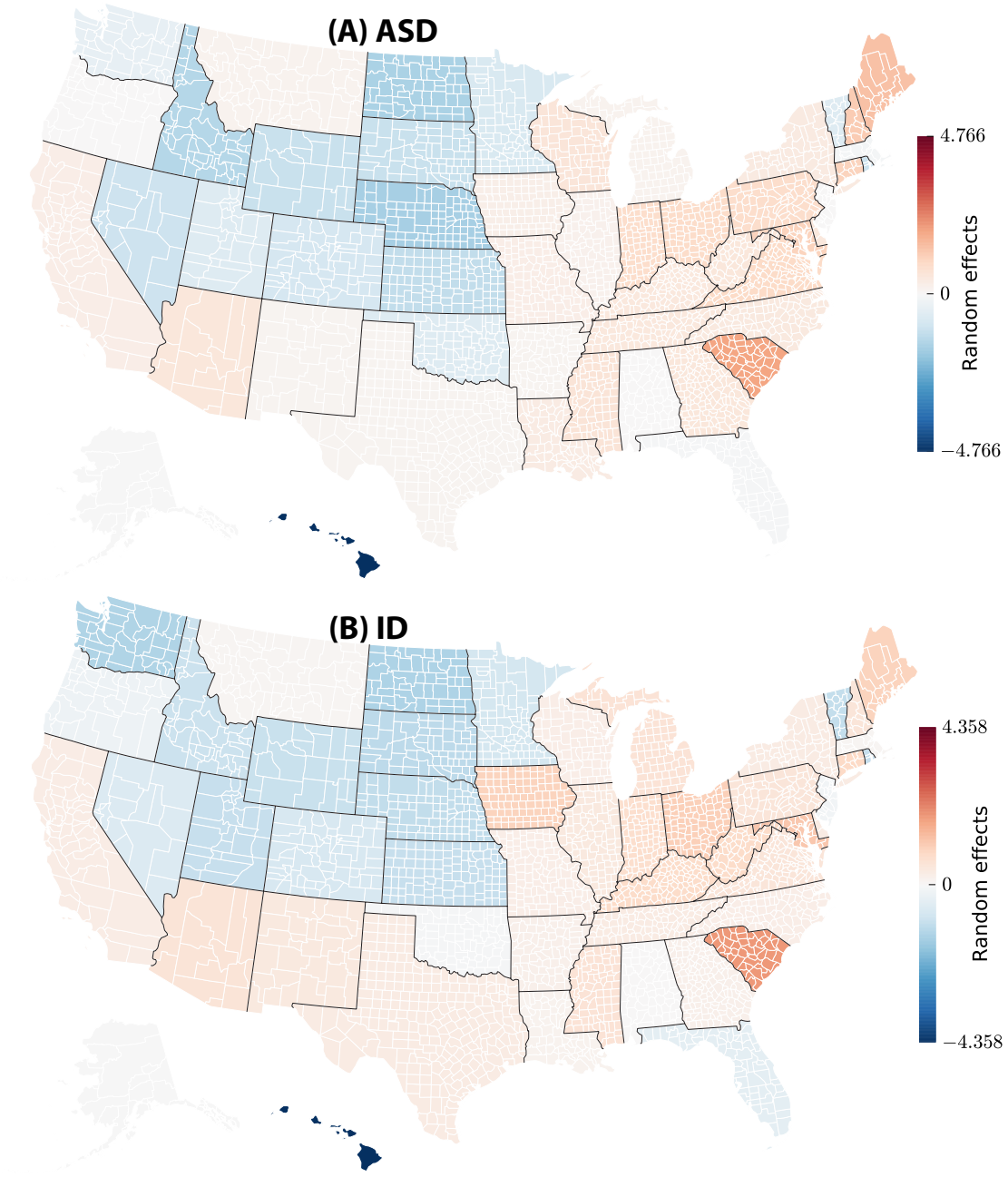


**Figure 2.5: Total county-level random effects of ASD and ID incidence in the USA: (A) ASD and (B) ID.** In the figures we color-coded the Empirical Bayes estimates of the state-level random effects, separately for ASD and ID. While county-specific random effects are directly comparable within the same state, comparison of these effects across different states is not meaningful, because each state-specific random effect determines the baseline disease rate for each county in the corresponding state, and these baseline rates vary across states.

of the two diseases significantly varied across ethnic groups (Table 1), with Pacific Islanders, for example, having significantly lower risk for both diseases. The per capita income of the county was weakly positively correlated with the incidence rates for both diseases: the income variable was associated with 3.2% rate increase per every additional $1,000 of income above the country average for ASD (CI: [2.3 %, 4.2%], $p < 6 \times 10^{-5}$) and a 2.7% rate increase for ID (CI: [1.8%, 3.7%], $p < 6 \times 10^{-5}$). Other important socioeconomic predictors included the percentage of urban population in a county; a one percent increase in urbanization predicted about a 3% in ASD and ID incidence (Table 1 A). Our analysis also indicated that state-specific laws [30,31] had a large but only marginally significant effect on the incidence rates of ASD and ID (Table 2.1A, Figure 2.1). The strictest form of diagnostic evaluation (variable Eval, the state-mandated diagnosis of autism or autism spectrum disorders by a pediatrician or clinician for consideration in the special education system) was predictive of a considerable decrease in ASD and ID incidence rates, 98.6% (CI: [28%, 99.99%], p=0.02475) and 99% (CI: [68%, 99.99%], p = 0.00637) respectively.

### *2.2.3 Discussion*

By analyzing the spatial incidence patterns of autism and intellectual disability drawn from insurance claims for nearly one third of the total US population, we found strong statistical evidence that environmental factors drive the apparent spatial heterogeneity of both phenotypes while economic incentives and population structure appear to have weaker effects. The strongest predictors for autism were associated with the environment: congenital malformations of the reproductive system in males (an increase in ASD incidence by 283% for every per cent of increase in the incidence of malformations), non-reproductive congenital malformations (31.8% ASD rate increase), and viral infections in males (19% ASD rate increase). For ID we observed similar but weaker effects: 93% increase of ID rate for every per cent of increase in congenital malformations of the reproductive system in males, 43% increase for per cent of non-reproductive congenital malformations, and 23% for viral infections in

males.

We highlight the role of male congenital genitourinary malformations as an approximate measurement of environmental exposure to unmeasured developmental risk factors, including toxins. Some infants are born with congenital malformations with unknown genetic etiology not explained by known Mendelian mutations or detectable chromosomal aberrations. At least a fraction of such birth defects may be due to parental exposure to environmental insults. The environmental factors implicated so far include pesticides [87, 89], environmental lead [249], sex hormone analogs [123, 122], medications [234], plasticizers [142], and other synthetic molecules [165]. More generally, the risk of congenital birth defects is statistically linked to parental occupation [139, 118, 170]. There is a statistically significant increase in birth defects associated with some maternal occupations (janitor, maid, landscaper), and significant decrease associated with others (non-preschool teacher) [118, 170]. It is very likely that the list of environmental factors potentially affecting development of human embryo is large and yet predominantly undocumented; correspondingly, detailed statistics on these factors do not exist.

It is known that some birth malformations are caused by de novo genetic events, such as large copy number variants that have been found to increase the risk for ASD by approximately 400% [11]. Single-gene deletions, for example, involving CHD7 are known to cause CHARGE syndrome [46, 219] associated with genital abnormalities and putatively associated with ASD [215]. However, these genetic events may have currently poorly identified environmental triggers, and 70 to 80% of male congenital malformations of the reproductive system have no clear genetic causes [16]. Instead, they appear to be driven by specific environmental insults that were not serious enough to lead to more serious adverse events during pregnancy, such as spontaneous abortion. Therefore, in this study, we used the rate of birth malformations as a surrogate measure for environmental burden.

The hypospadias of the male urethra can arise during early embryonic development, specifically weeks 9-12 (p. 206 in [16]). This window corresponds to the time when cell

division and migration takes place in brain development. Furthermore, maternal exposure to estrogen and estrogen analogs in animal models affects both brain and genital development in male progeny (p. 206 in [16]), and small physical malformations appear enriched in autistic children compared to healthy children [218].

Following similar logic, in addition to causing birth defects, environmental toxins, such as pesticides [235, 195] can substantially weaken the human immune system, especially in men, which results in more frequent infections. (The rates of female viral infections were highly correlated with male viral infections; these can serve a somewhat weaker fixed effect predictor, data not shown.) This suggests that per capita rate of viral infection, when socioeconomic and other biological factors have been controlled for, may serve as another environmental indicator, although specifics of the causal, biological mechanisms remain unresolved. In our analysis we found that the rate of viral infection in males was significant for both ASD and ID, see Table 1.

Importantly, the effect of state-level regulations involving ASD appeared relatively large in magnitude (over 98% ASD and ID rate decrease) but with a wide confidence interval and inconsistent effects across states, resulting in only marginal significance. Furthermore, our estimates of random effects at the state and county levels, see Figure 2.2, suggest that additional yet unknown confounder factors exist at both state and county levels, as is evident from the clear state boundaries seen in Figure 2.2.

As with other statistical analyses, significant associations are not necessarily causal. Identified predictor variables may reflect underlying mechanisms, or may be correlated with unmeasured causal factors. However, we have included variables, such as the rate of birth malformations and male viral infections, that have well-documented environmental causes. Overall, this increases our confidence in their scientific relevance. Furthermore, because we have controlled for many county-level socioeconomic variables, strong state-specific effects are almost certainly rooted in legal and regulatory differences that exist at this level. Our results have implications for the ongoing scientific quest for the etiology of neurodevelop-

mental disorders. We provide evidence that routinely expanding the scope of inquiry to include environmental, demographic and socioeconomic factors, and governmental policies at a broad scale in a unified geospatial framework. It appears that detailed documentation of environmental factors should be recorded and used in genetic analyses of ASDs and failure to do so risks omitting important information about possibly strong confounders.

### *2.2.4 Methods*

Our multi-level, mixed-effects model predicted the incidence of ASD and ID conditional on several individual-level, county-level, and state-level covariates. For the regression analysis described below, we used county level variables to predict disease rate. In the analysis of the comorbidity between congenital malformations and ASD or ID, we used patient-level data.

Data: We used the Truven Health Analytics MarketScan® Commercial Claims and Encounters Database to provide geocoded diagnosis counts by gender. This database spans the years 2003 to 2010 and consists of approved commercial health insurance claims for between 17.5 and 45.2 million people annually, with linkage across years, yielding a total of approximately 105 million patient records. (Note that, consistent with low prevalence of both phenotypes, only a small proportion of individuals described in this enormous dataset were diagnosed with either ASD or ID.) This national database contains information contributed by well over 100 insurance carriers and large self-insuring companies. We scanned the approximately 4.6 billion inpatient and outpatient service claims and identified almost 6 billion diagnostic codes. After removing duplicates, almost 1.3 billion diagnostic codes were found to be associated with over 99.1 million individuals, yielding approximately 12.89 unique diagnostic codes per individual. Claims were de-identified, that is, all patient-level personal information was redacted, and geocoded at the county level by Truven, and thus, this did not require any additional processing on the authors' part.

The MarketScan insurance claims dataset is not a truly random sample of the USA population. This is because compilation of this dataset required reaching agreements be-

tween Truven and numerous individual insurance providers to share data, and the insurance providers inherently had uneven and non-random coverage of geographic areas. It is possible, therefore, that the dataset carries traces of hidden correlations imposed by the data collection method. Furthermore, while the entire USA is well represented in the data, it is possible that coverage across geographic areas is not perfectly proportional to population density.

Statistical Analysis: We framed our analysis as a mixed-effect regression model for Poisson-distributed count data [114], independently implemented in SuperMix [115], lme4 [280, 18], MCMCglmm [107], and GLLAMM [270, 232]. The choice of the Poisson model was motivated by the countable nature of data and the rarity of the disease incidence events.

The model parameters were estimated using a joint statistical inference as follows. Most of the parameters (44 out of 50) were real-valued coefficients representing regression weights of individual factors, such as average income in county, percentage of ethnic groups, per cent of urban and poor population, see Table 2.1. The factors that are a priori suspected to be relevant to disease incidence and are deliberately included into the model, are referred to in the mixed effect model formalism as the fixed effects. In addition, the model included zero-centered and normally distributed random effects, uncorrelated for the same disease between a county and the encapsulating state, but geographically correlated between two diseases, see equations below and Figures 2.4 and 2.5. The six parameters associated with the random effects included the variances and covariances for the state- and county-level random effects (see Table 2.2 and equations below). Note that the random effects themselves were not parameters, but Gaussian zero-centered random variables.

Fixed effects by design have stochastic but predictable influence on data, while random effects describe zero-centered random influence, not captured by the fixed effects. Below we present more formal, formulation of the model.

Fixed-Effect Variables: The Truven database was augmented with US census data [112] consisting of county-level measurements for a variety of socioeconomic, demographic, and

geospatial factors. Our fixed-effect county-level covariates were gender, Gender, average per capita income, Income, percent ethnicity (separately for American Indians, AmInd, Asians, Asian, White Hispanics, WHisp, White non-Hispanics, W, Black Hispanics, BHisp, Black non-Hispanics, B, and Pacific Islanders, Pacific), and the proportions of various socioeconomic groups (poor, Poor, urban, Urban, insured, Insured). Our county-level environmental indicators used as fixed-effect covariates (normalized by county population size) comprised congenital malformations excluding malformations of the genitals (separately for females and males, CongMrepF and CongMrepM, respectively), congenital malformations of the genitals (separately for females and males, ConGenF and ConGenM, respectively), viral infections (separately for females and males, Viral_F and Viral_M, respectively), ectopic pregnancy (ect_pr), abnormal conception (abnormal_concept), spontaneous abortion(spont_abort), and multiple gestations (mult_gest). The county-level environmental indicators were extracted from the Truven database and normalized by county population, separately for males and females.

To account for variation in policies for special education eligibility and reimbursement, we used four variables derived by hand-coding the state policies for eligibility in special education programs under the Individuals with Disabilities Education Act [182] with categorical variables: (i) CFR, to indicate whether state criteria met (-1) or, alternatively, exceeded Code of Federal Regulation requirements (1), (ii) DSM, to indicate whether state criteria mentioned all of the criteria from the Diagnostic and Statistical Manual of Mental Disorders (-1, if no, and 1, if yes), (iii) ASD, to indicate whether Autism Spectrum Disorder criteria were mentioned in the state criteria (-1 if no, 1 if yes), and (iv) Eval, to indicate the degree of diagnostic rigor required by the state (-1, -2, 1, 2). Details of the coding are described in the abbreviations. All predictor variables were mean-centered.

Assumptions of the model The assumptions of the Poisson regression model [114] were as follows. First, we assumed that the data, corresponding to the observed counts of people within each county diagnosed with either ASD or ID, were generated by a Poisson process,

with rate $(ijkl)$ varying over counties,

$$f(y_{ijkl}|\theta) = \frac{\exp(-\lambda_{ijkl})\lambda_{ijkl}^{y_{ijkl}}}{y_{ijkl}!}$$

where: is a vector of all 50 model parameters, (b, ). The observed counts of disease incidence (the response variable $y^{ijkl}$) was defined as the number of disease cases per county for ASD (k=1) and ID (k=2). Subscripts i and ij were used to indicate a state and a county nested within that state, and subscript l to indicate gender. The second assumption was that the logarithm of Poisson rate $(ijkl)$ was expressed as a linear combination of fixed and random effects.

$$\begin{aligned}\lambda_{ijkl} =& N_{ijkl} \exp(\mathbf{X}_{ik}\mathbf{b}_k + \mathbf{z}_k\mathbf{v}_k) \\ =& N_{ijkl} \exp\left(\Sigma_{m=1}^{M} x_{imk}\beta_m k + v_{ik} + v_{ijk}\right)\end{aligned}$$

Here matrix $X_k$ is the design matrix for the fixed effects associated with disease k; bk is the corresponding vector of unknown regression weights; zk is a design matrix for random effects; $v_k$ is the vector of random effects; $v_{ik}$ , vijk, are i-state- and ij-county-specific random effects for disease k. The fixed-effect design matrix is simply a matrix of county-specific zero-centered properties, such as the mean income, or proportions of ethnic groups. The design matrix z has a very simple form: entries of 1 for random effects of a given county and corresponding state, and zeros in all other cells. $N_{ijkl}$ is a state-, county-, disease- and gender-specific offset—the total number of people with a specified gender living within a given county.

The third assumption was that data was hierarchical: the zero-centered random effects were independently introduced at i-state and ij-county levels. However, the random effects

associated with the two diseases (1 and 2) were geographically correlated,

$$
\begin{aligned}
(v_{i1}, v_{i2}, v_{ij1}, v_{ij2}) \sim & \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{i1}^2 & r_i & \sigma_{i1}\sigma_{i2} & 0 & 0 \\ r_i\sigma_{i1}\sigma_{i2} & \sigma_{i2}^2 & 0 & 0 & \\ 0 & 0 & \sigma_{ij1}^2 & r_i\sigma_{ij2}\sigma_{ij2}^2 & \\ 0 & 0 & r_{ij}\sigma_{ij1}\sigma_{ij2} & \sigma_{ij2}^2 & 0 \quad 0 \end{pmatrix} \right) \\
= & \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})
\end{aligned}
$$

where $2_ik$, and $2_ijk$, are variances of state-level and county-level random effects for disease $k$ and $r_i$, and $r_{ij}$ are correlation coefficients for random effects for diseases 1 and 2 at the state and county levels, correspondingly.

Together these assumptions define the following likelihood for the given ij-county, i-state, l-gender and k-disease:

The full log-likelihood is obtained by summing the individual log-likelihoods specific to each $y_ijkl$ over all possible indices.

$$
\begin{aligned}
\log l(y_{ijkl}|\theta) = & \big[ - N_{ijkl} \exp(\Sigma_{m=1}^{M} x_{imk}\beta_m k + v_{ik} + v_{ijk}) \\
& + y_{ijkl}(\log N_{ijkl} + \Sigma_{m=1}^{M} x_{imk}\beta_{mk} + v_{ik} + v_{ijk}) - \log(y_{ijkl}!)\big]
\end{aligned}
$$

While estimates varied slightly across different implementations of the model and estimation approaches, the major trends were identical across all. Here we present the results of the Markov chain Monte Carlo/Empirical Bayes analysis. The estimation methods and starting parameter values varied considerably across the implementations. For example, SuperMix started with finding an analytical solution of the fixed-effect part of the equations and then estimated parameters for the full model involving random effects. The Markov chain Monte Carlo-based GLLAMM started with a random set of parameter guesses and then rather quickly discovered the high-probability area of parameter values.

Confounders We tested several putative confounding variables, such as county-specific

median mother's age at childbirth, and the proportion of county population in the child-bearing age. While these putative confounding variables were associated with statistically significant fixed-effect coefficients, they did not affect the relationship between the outcome variable and the compound environmental predictor variables.

## 2.3 Understanding Toxoplasmosis in the United States through "Large Data" Analyses

### *2.3.1 Introduction*

Toxoplasmosis, disease caused by infection with the Apicomplexan parasite, Toxoplasma gondii, is a major source of morbidity and mortality in the United States and globally [86]. Disease presentation is highly variable, though severe eye disease resulting in loss of sight is not uncommon, and immune-compromised patients can present with serious infections of the central nervous system [162, 8, 314, 102, 60, 319]. Acquisition of infection during pregnancy can result in vertical transmission, leading to profound disability due to untreated congenital infection; consequences for the infected, untreated infant include cognitive impairment, hydrocephalus, and disability due to loss of sight [189]. Additionally, previous studies posit a role for *T. gondii* infection in a variety of comorbid conditions, including epilepsy and neurologic diseases [2, 31, 287, 134, 192]. In these ways, toxoplasmosis continues to represent a major public health threat, as yet incompletely characterized in the U.S. Thorough quantitation and characterization would facilitate targeted intervention strategies. To our knowledge, there are no accurate, empiric data defining incidence of toxoplasmosis in the U.S.. According to the CDC, toxoplasmosis remains one of the leading single causes of death related to food-borne illness in the U.S., although numerically it is a minority. It is a neglected infection [37]. Estimates of infection prevalence are limited in scope. Seroprevalence in the U.S. is estimated at 11% for women of child bearing age[146]. Additionally, one recent study estimates 4,839 cases of symptomatic ocular toxoplasmosis per year nationally.

[145]. Estimates for other manifestations, including infection of the central nervous system (CNS) and visceral organs, are lacking. Most studies of prevalence of toxoplasmic meningoencephalitis in the U.S. were from the 1990s, in the context of the HIV/AIDS epidemic, prior to advent of HAART [228, 144]. The most recent estimate, in 2003, indicated an annual risk reduction of 18% for cerebral toxoplasmosis, though these data were derived from a cohort of individuals being treated with HAART [250]. There is a paucity of estimates of non-eye manifestations for the last decade. Active infection with *T. gondii* is treated successfully with pyrimethamine and sulfadiazine [60, 319, 189, 126]. These are the most effective medicines for treating toxoplasmosis, including ocular and CNS manifestations, and are not indicated in other clinical contexts. Additionally, there has been increasing use of trimethoprim-sulfamethoxazole (TMP-SMX) to treat ocular toxoplasmosis, with some authors suggesting similar clinical outcomes, although TMP-SMX is ten fold less active with suboptimal ratios of constituent medications with similar risks of hypersensitivity [286, 263]. Given gaps in knowledge of incidence in the U.S., and uncertainties about frequency of different disease manifestations and other potential comorbid conditions, we use a novel approach to understand epidemiology of toxoplasmosis, namely use of large insurance-based datasets. Truven Health MarketScan® Commercial Claims and Encounter Database was queried to answer questions related to epidemiology of this infection. This database contains insurance claims from privately insured patients, not including individuals with access only to Medicare or Medicaid or without any health insurance. This database represents approximately 15% of the total U.S. population. The MarketScan® Databases have been used many times in the context of infectious diseases epidemiology and cost-related studies, although we did not find such reports for toxoplasmosis [216, 49, 217, 117, 82, 90]. Use of large datasets to characterize epidemiology of infectious disease presents a unique opportunity to assess prevalence and incidence of infection with minimal cost, while allowing for characterization of occurrence across time and space. Additionally, this approach facilitates identification of patterns of infection with respect to patient age, gender, and common comorbidities. Herein,

we present a novel approach to understanding epidemiology of toxoplasmosis through use of "large data."

### *2.3.2 Results*

Study Population ICD-9 codes specifically indicating infection with *T. gondii* identified 9260 unique patients from 2003-2012, out of a total of 151 million patients. Use of medicine codes for pyrimethamine or sulfadiazine, and TMP-SMX in context of *T. gondii* infection or eye disease, identified 2305 and 7690 cases, respectively. There was an overlap of 225 cases between those with toxoplasmosis-specific drugs and cases where TMP-SMX was used. Disease Manifestations Toxoplasmic chorioretinitis accounted for 3,492 (37.7%) identified patients. Meningoencephalitis occurred in 472 (5.1%) patients. Less common manifestations of infection include myocarditis (113), pneumonitis (113), and hepatitis (188). These conditions comprised 1.2 % of total cases each for myocarditis and pneumonitis, and 2.0% of total cases for hepatitis. Disseminated toxoplasmosis occurred in 120 patients. Codes indicating unspecified toxoplasmosis were assigned to 4194 (45.3%) patients.

Demographics Patients ranged in age from 0 to 70 years. 218 (2.4%) occurred in children two years of age and younger, indicating likely congenital infection. Mean age at diagnosis was 37.5 $\pm$ 15.5 years. Approximately 41% (3,776) of cases occurred in males, while the remaining 5,484 occurred in females. One demographic of particular interest is women of reproductive age, here defined as between age 13 and 51 years, who have potential for vertical transmission to a fetus. In this cohort, 73% (4,022) were of reproductive age.

Geographic Distribution of Toxoplasmosis Cases 9,260 identified cases were divided up based on U.S. Census region. South, including South Atlantic and East and West South Central census regions, encompassed most identified cases, with 4,614/9,260, or almost 50% of cases. West, including Mountain and Pacific regions, contained only 1,205/9,260 cases, or 13%. $\chi^2$ analysis revealed distribution of cases was not consistent with distribution based on population, p ¡ 0.001. See Figure figs. 2.7 and 2.8 for distribution of cases across census

regions and a map of disease prevalence across the country.

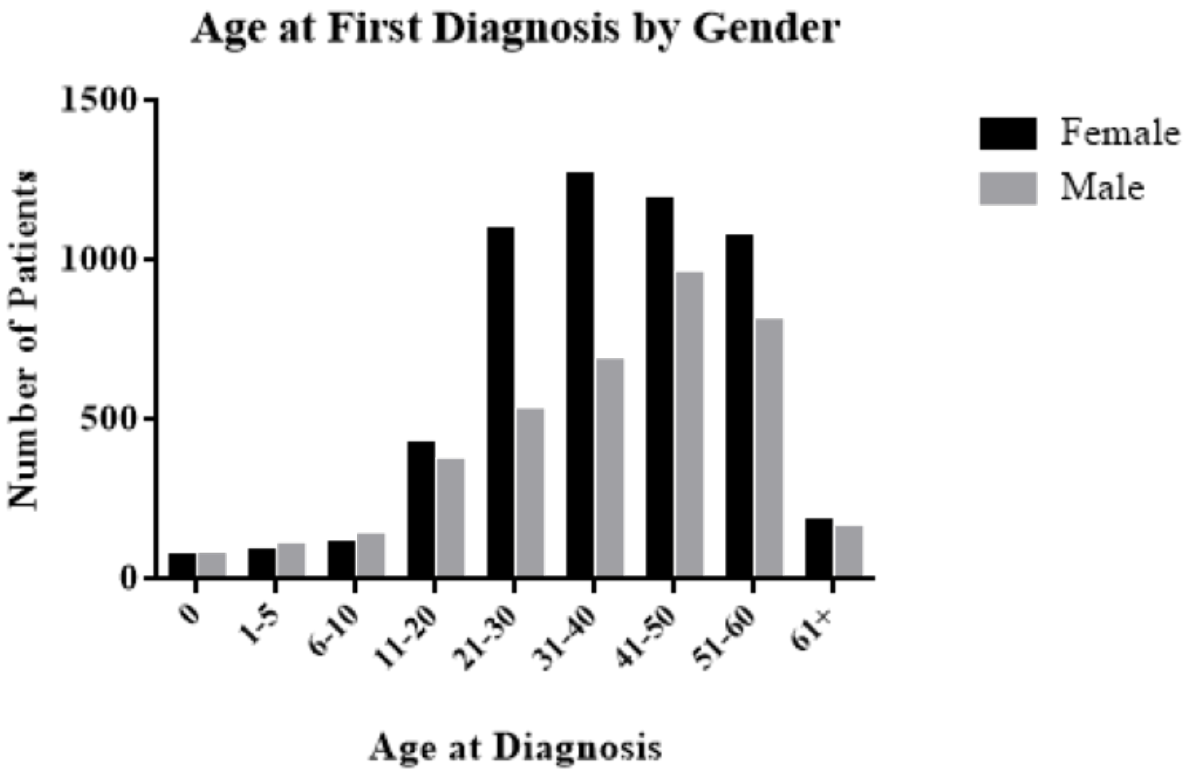


**Figure 2.6: Age at first diagnosis with toxoplasmosis.** Increasing numbers of patients were diagnosed with toxoplasmosis as a function of age, indicating that there is an increasing disease burden among older individuals, who are potentially more capable of moving through the environment and engaging in risky behaviors, including the consumption of undercooked food, which pose the threat of exposure to infectious oocysts and ingestion of contaminated materials. More females than males were identified with toxoplasmosis in this cohorts (approximately 59% vs 41% of the total number of cases). The ratio of male to female was most similar early on in the patient's lifetime, while there was an increasing disparity skewing towards increased incidence in females. This could be explained by screening of women of reproductive age, or could potentially indicate a difference in risk of exposure to the pathogen.

Estimated Annual Case Burden by Disease Manifestation in the U.S. This study identified 9,260 cases of toxoplasmosis over the period from 2003-2012, corresponding to 61,373 cases over the study period, or a rate of 6,137 cases per year for the whole population of the United States. Estimated annual incidence by disease manifestation is listed in Table 2.4.

Comorbidity in *T. gondii* Infection based on specific Toxoplasma ICD 9 code. Odds

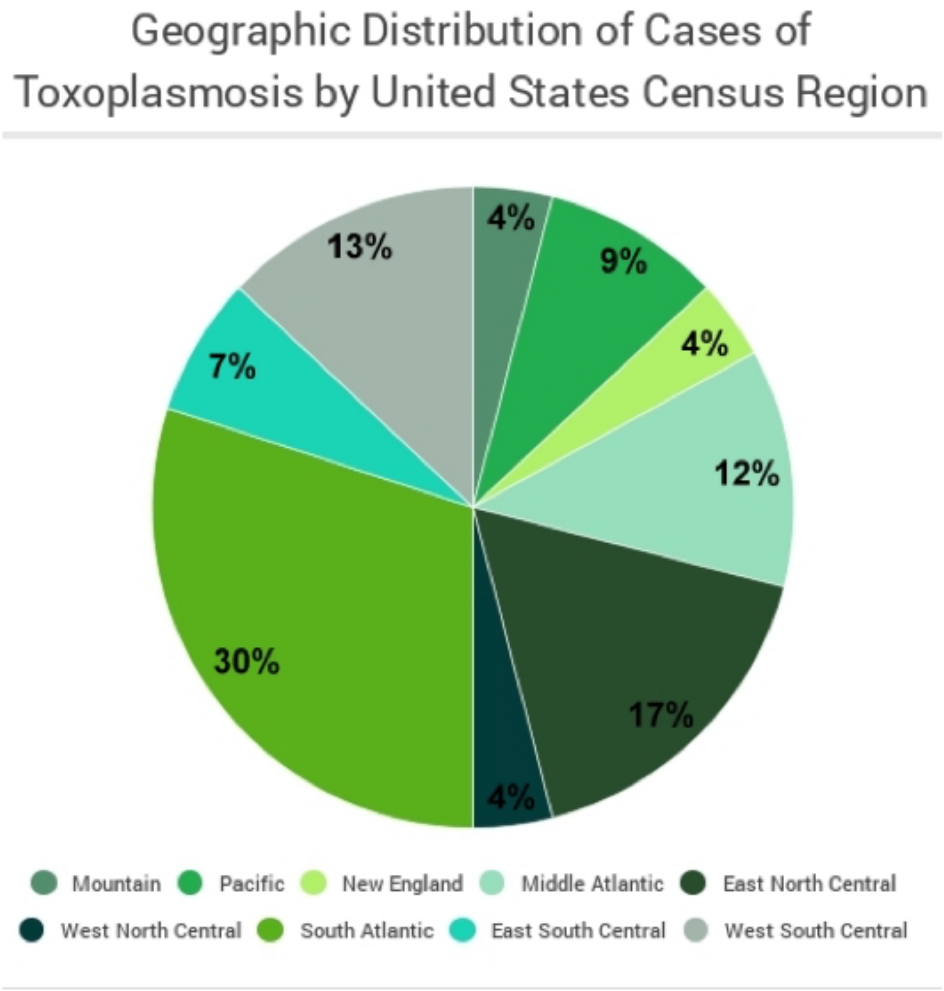


**Figure 2.7: Pie chart showing geographic distribution of toxoplasmosis cases.** More cases of toxoplasmosis occurred in the southern U.S., which is consistent with enhanced persistence of oocysts in the environment in warm and wet climates. There is also substantial agricultural activity in this region of the country, which could also explain increased risk of exposure

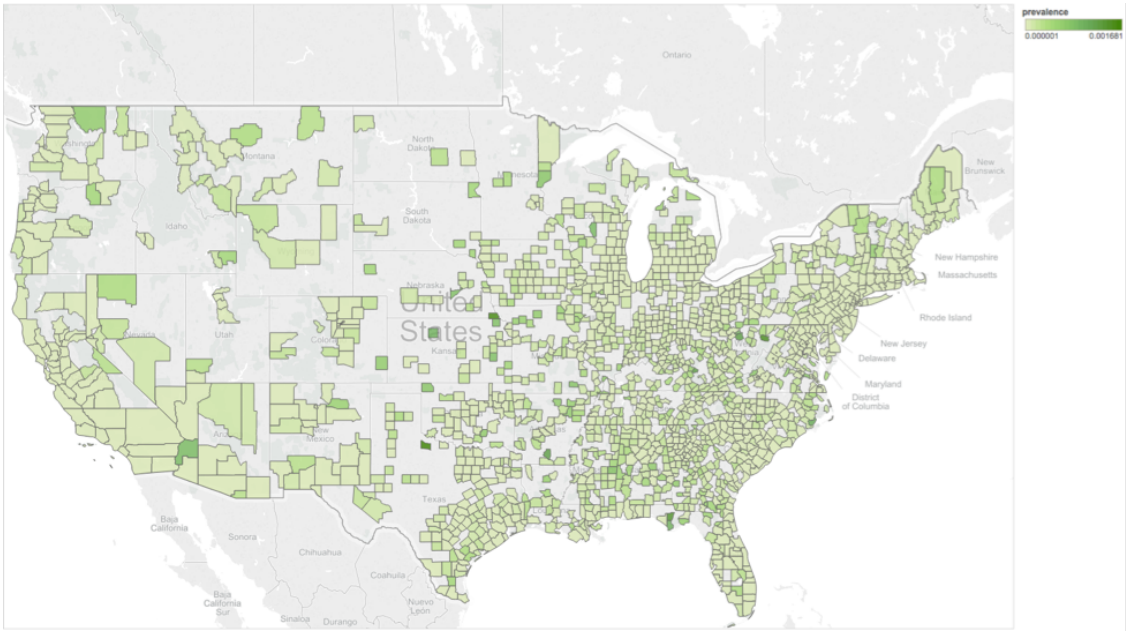


**Figure 2.8: Map of toxoplasmosis prevalence by county.** Darker regions indicate increased prevalence of the infection. States with highest prevalence include Texas, California, New York, Illinois, Georgia, Florida, Maryland, and South Carolina. All these states had more than 400 patients with toxoplasmosis over the study period. There are pockets of increased prevalence across the southern U.S. This is consistent with previous studies indicating increased prevalence among rural populations, with increased risk of environmental exposure

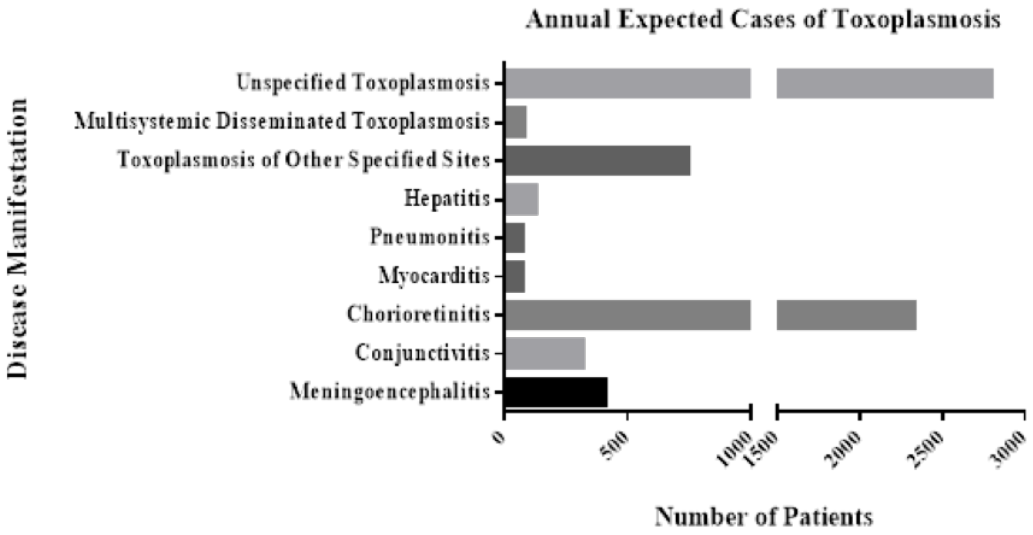


**Figure 2.9: Clinical Manifestations of Toxoplasmosis, 2003-2012.** A total of 9260 individual patients with toxoplasmosis were identified. The majority of cases (4194) were coded as unspecified toxoplasmosis. Eye disease due to toxoplasmosis, encoded as chorioretinitis and conjunctivitis, accounted for 3492 and 472 cases respectively. It is important to note that toxoplasmic conjunctivitis is not a known clinical condition, and it is likely that this was simply coded for toxoplasmic eye disease, more generally. Toxoplasmosis of other specified sites was coded in 1114 patients. 610 instances of patients with CNS infection due to *T. gondii* were recorded. Liver infection accounted for 188 cases, while infection of the lung and myocardium both accounted for 113. Multisystemic disseminated toxoplasmosis, with diffuse organ involvement, was identified in 120 patients.

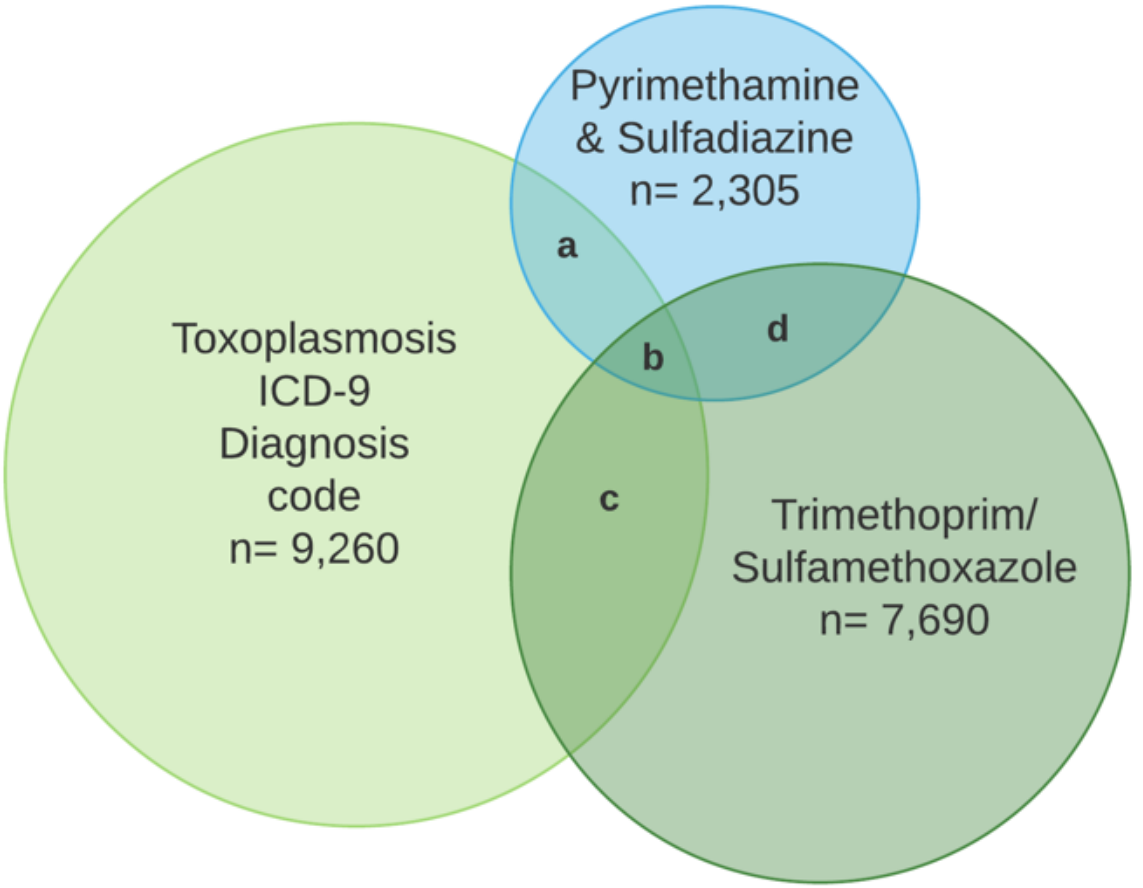


**Figure 2.10: Venn diagram showing numbers of patients who were diagnosed by ICD 9 codes for Toxoplasmosis with and without codes for medicines.**

ratios for statistically significant comorbidities of interest in patients with codes indicating toxoplasmosis, compared to matched controls without these codes, can be found in Table 2.5.

| Disease Manifestation | Number of Cases Identified by MarketScan® Database, 2003-2012 | Estimated Number of Cases in U.S., 2003-2012 | Estimated Annual Incidence |
|---|---|---|---|
| Meningoencephalitis | 610 | 4,067 | 407 |
| Conjunctivitis | 472 | 3,147 | 315 |
| Chorioretinitis | 3,492 | 23,280 | 2,328 |
| Myocarditis | 113 | 753 | 75 |
| Pneumonitis | 113 | 753 | 75 |
| Hepatitis | 188 | 1,253 | 125 |
| Toxoplasmosis of Other Specified Sites | 1,114 | 7,427 | 743 |
| Multisystemic Disseminated | 120 | 800 | 80 |
| Unspecified Toxoplasmosis | 4,194 | 27,960 | 2796 |
| Total Number of Cases | 9,206 | 61,373 | 6,13 |

**Table 2.4: Number of Cases of Toxoplasmosis in the United States, by Disease Manifestation, and Estimated Annual Incidence.**

National Drug Code (NDC) Use in the Identification of Toxoplasmosis Cases In addition to patients with a diagnosis of toxoplasmosis coded, analyzed as above, patients were identified with NDCs, considered separately because diagnosis was less certain (Figure 2.10). Pyrimethamine and sulfadiazine are specific to treatment of infection with this parasite, indicating potential utility in identifying patients with active toxoplasmosis. However, treatment could have been presumptive and then discontinued when another disease was diagnosed. TMP-SMX sometimes has been used for treating retinal disease due to Toxoplasma, although it is a suboptimal treatment. Therefore, presence of eye disease and TMP-SMX treatment might suggest, but not confirm, this diagnosis. Other illnesses causing eye disease, and even different diseases requiring treatment with TMP-SMX, could confound this surrogate

| Comorbidity | OR Mean (95% CI) |
|---|---|
| HIV | 17.57 (14.61-21.13) |
| Malignant Brain Neoplasm | 8.69 (4.60-16.41) |
| Unspecified Encephalopathy (Hydrocephalus) | 5.55 (4.75-6.48) |
| Epilepsy | 3.51 (3.00-4.12) |
| Thrombocytopenia | 3.17 (2.64-3.80) |
| Benign Brain Neoplasm | 2.80 (2.08-3.74) |
| Visual Loss (Acquired Visual Disturbances) | 2.55 (2.39-2.73) |
| Systemic Lupus Erythematosus | 2.37 (1.88-2.99) |
| Schizophrenia | 2.21 (1.80-2.72) |
| Multiple Sclerosis | 2.05 (1.64-2.58) |
| Gestational and Pregnancy Related Disorder | 1.92 (1.80-2.05) |
| IBS and Crohn's Disease (Regional Enteritis, Crohn's Disease) | 1.49 (1.33-1.67) |
| Bipolar Disorder | 1.38 (1.17-1.62) |
| Substance Abuse | 1.24 (1.14-1.36) |
| Anxiety | 1.24 (1.16-1.33) |

**Table 2.5: Comorbidity Odds Ratios with toxoplasmosis ICD-9 Codes.**

marker for active toxoplasmosis. As shown in Figure 2.10, there were 2080 patients with only pyrimethamine and/or sulfadiazine; 7465 patients with only TMP-SMX plus chorioretinitis. In addition , 225 patients took both, not necessarily simultaneously. Of the total 9260 patients with a toxoplasmosis diagnostic code, 339 also had a pyrimethamine and/or sulfadiazine code; 690 also had a TMP-SMX code; and 152 had both. NDC cohort demographics, and a comparison between the two medicine treatment groups were similar to the toxoplasmosis ICD-9 code group, except age was younger ($\bar{x}$= 38 years [ICD9 toxoplasmosis] vs 51 years [NDC]). Also, there were more patients in middle and south Atlantic states in the toxoplasmosis code group. Odds ratios for co-morbidities comparing treatment with pyrimethamine/sulfadiazine versus TMP-SMX-chorioretinitis group are shown in Table 2.5b and are similar to those with ICD-9 codes for toxoplasmosis.

### *2.3.3 Discussion*

Use of Truven Health MarketScan® Commercial Claims and Encounter Database presented a novel opportunity to assess prevalence of medical diagnosis of toxoplasmosis in the U.S. population with indemnity insurance, confirmed some trends highlighted earlier, and yielded some unexpected results. Almost half of identified patients had codes indicating "unspecified toxoplasmosis," making it impossible to accurately quantitate disease manifestations. Of cases of toxoplasmosis with specified clinical manifestation, toxoplasmic chorioretinitis represented the highest proportion. This indicates a comparatively high frequency of eye disease relative to other manifestations, that eye disease due to *T. gondii* is more often diagnosed, or that it drives patients to seek care more than occurs for patients with other clinical manifestations. Our data do not provide support for other explanations for this observation. Almost 500 patients presented with one of the most lethal complications of toxoplasmosis, meningoencephalitis. This emphasizes that immune-compromised patients remain at risk for life-threatening manifestations. Our estimate indicates an annual disease burden of 488 cases of CNS disease due to *T. gondii*. Other, less common, manifestations of toxoplasmosis, including myocarditis, pneumonitis, and hepatitis cause morbidity and mortality in the U.S. It is critical that the index of suspicion for these rarer manifestations remain high, especially in immune-compromised patients, and that appropriate therapy is initiated when toxoplasmosis is suspected. Numbers of cases of toxoplasmosis, identified by ICD-9 code, were higher in the southern U.S. than would be expected based on population. The environment is more conducive to extended viability of the environmentally resistant, highly infective oocyst stage in this region [166]. Additionally, behavioral factors like agricultural activity may represent an additional explanation for this comparatively high prevalence. Moreover, the western census region demonstrated fewer cases than expected. An environment that is dryer and more hostile to oocysts, or differences in parasite or vector distribution, or recognition and reporting could explain this. Prevalence of comorbidities associated with toxoplasmosis was also assessed, affording insight into how commonly individuals with this infection suffer from

other conditions. Compared to controls matched for age, geography, and health, patients with codes for toxoplasmosis had greater odds of suffering from conditions including HIV, benign and malignant brain neoplasm, epilepsy, autoimmune diseases including lupus and multiple sclerosis, and psychiatric conditions including substance abuse, anxiety, bipolar disorder, and schizophrenia. Directionality and causality of these relationships is not clear. However, immunosuppression, with HIV infection or with immunosuppressive therapies such as lupus or malignancy, predisposes patients to severe infection. Chronic inflammation produced by presence of organisms in brain may promote neoplastic transformation and epilepsy, as could focal lesions or changes in neurotransmitters as occurs in animal models [100]. Literature has repeatedly demonstrated association between seroprevalence of *T. gondii* and neuropsychiatric illness, and literature has not demonstrated directionality or causality for these associations. Ascertainment bias could contribute to odds ratios, with neurologic signs and symptoms precipitating testing for toxoplasmosis. There are certainly billions of infected persons without neurobehavioral disease, so if there is a real association, other factors, such as host or parasite genetics or timing of infection, must be at play. Furthermore, since this cohort is predominantly under the age of 65 years of age and without public insurance, it remains of interest to study the effect of this possible source of chronic inflammation in the brain on neurodegenerative diseases in the geriatric population.

While this analysis has offered potentially valuable insight into toxoplasmosis in the U.S., the use of insurance databases for epidemiology has obvious limitations which must be addressed. Disease is a complicated phenomenon, with many factors that influence its prevalence and distribution. Host behavior can influence rates of parasite transmission and risk of acquisition. This database comprises patients that are privately insured, and does not include Medicaid or Medicare patients, or those without insurance. This introduces sampling bias. While *T. gondii* is an equal opportunity parasite, capable of infecting people irrespective of socioeconomic status, there are risk factors that make infection more likely in certain populations, especially southern agricultural laborers, where a lack of private health

insurance is not uncommon. In the calculations presented, we assume that the population in the database is reflective of the population as a whole, even though it is likely this is not the case, and thus our analysis likely underestimates actual rates of infection and disease due to *T. gondii*. Anecdotally, of 8 patients seen by our group in a private hospital setting with toxoplasmosis in the last month, 6 did not have private insurance, and therefore would not have been identified by this approach. If this is reflective of the general population, there is substantial underestimation using this approach. Thus it would require a more aggressive public health approach to toxoplasmosis or another method for estimating prevalence.

Additionally, the way in which physicians use ICD-9 codes in the context of toxoplasmosis presents a limitation to assessment. Most patients whose claims contain references directly to toxoplasmosis have codes indicating "unspecified toxoplasmosis," which offers no information about symptoms. As such, this approach to characterizing toxoplasmosis on a population level is unlikely to be reflective of the true diversity of disease manifestation, which can range from lymphadenopathy, which is less likely to drive a patient to seek care when self-limited, to meningoencephalitis, a potentially fatal form of infection. Of note, more individuals had codes indicating treatment for infection with *T. gondii* than had claims indicating the infection. Approximately 6% of patients did not have ICD-9 codes for infection, even though they received medication used to treat this infection. There is the possibility that a substantial percentage of patients are not being coded appropriately when infected. Thus, more people are receiving medications than would be detected by ICD-9 codes, indicating a limitation of using ICD-9 codes for epidemiological studies. If previous estimates of disease burden are correct, this method underestimates prevalence. Annual burden of symptomatic eye disease due to infection with *T. gondii* has been estimated at 4,839 cases annually (16). Our analysis conservatively estimates 2,643 cases per year. Thus, this analysis underestimates cases by almost half, if previous reports are accurate. Other manifestations of toxoplasmosis, including meningoencephalitis, have not been estimated in the years since the development of HAART therapy, limiting comparative analyses with the

MarketScan® Database for CNS infection. Congenital infection remains a substantial cause of significant human suffering in association with vertical transmission from mother to fetus. Of the identified patients, 228 were between the ages of 0 and 2 at diagnosis. Detection of this disease at birth identifies significantly symptomatic patients, which would be expected to represent only a small proportion of these infections. Most would be mild and not come to medical attention in the absence of systematic screening. Forty four percent of patients were women of reproductive age indicating that this infection likely poses considerable medical burden during fetal life. A number of countries, including France, have implemented mandatory screening during gestation, for infection with *T. gondii*. This approach has been demonstrated to be efficacious in reducing frequency and severity of congenital infection by facilitating early diagnosis and treatment, and to be cost-beneficial [238, 274, 246, 302]. Screening would facilitate appropriate treatment of acutely infected mothers, which has been demonstrated to reduce disease severity in and frequency of infected infants.

Our observations of inability of analyses of large databases like MarketScan® CCAE to identify all cases of toxoplasmosis in the U.S. presents a different type of opportunity for intervention. The fact that physicians are not coding for toxoplasmosis, either due to an inability to recognize it, a lack of time to code for it, or some other barrier, in addition to the fact that this parasite presents a health threat to the American population as well as globally, suggests that it may be useful to make reporting of the disease to health departments mandatory. This would facilitate more accurate assessment of prevalence and severity of toxoplasmosis in the U.S., which could enable public health interventions that will ultimately reduce human suffering and mortality. Toxoplasmosis is a treatable condition, but an inability to appreciate the magnitude of the problem in this country presents a barrier to appropriate management of disease due to this parasite.

### *2.3.4 Methods*

Database Information Truven Health MarketScan Commercial Claims and Encounter Database includes privately insured patients, and was used to identify individuals with toxoplasmosis from 2003-2012. (See supplemental).

Identification of Patients and Assessment of Demographic Information 130.x series of ICD-9 codes (Table 2.6) are used to specifically indicate infection with *T. gondii*, including diverse manifestations, such as CNS infection and eye disease. To identify cases of toxoplasmosis not specifically coded for by ICD-9 codes, numbers of claims for medicines specific to treatment of toxoplasmosis also were assessed. The database identified patients who received these medicines, even in absence of an ICD-9 code indicating presence of this specific infection. Claims were considered to represent toxoplasmosis if they included a relevant ICD-9 code and a medicine use claim within 7 days of each other. Once patients with disease due to *T. gondii* were identified, patients were evaluated for age at first diagnosis, gender, and U.S. census region.

| Condition | ICD-9 Code |
|---|---|
| Meningoencephalitis due to Toxoplasmosis | 130.0 |
| Conjunctivitis due to Toxoplasmosis* | 130.1 |
| Chorioretinitis due to Toxoplasmosis | 130.2 |
| Myocarditis due to Toxoplasmosis | 130.3 |
| Pneumonitis due to Toxoplasmosis | 130.4 |
| Hepatitis due to Toxoplasmosis | 130.5 |
| Toxoplasmosis of Other Specified Sites | 130.7 |
| Multisystemic Disseminated Toxoplasmosis | 130.8 |
| Toxoplasmosis Unspecified | 130.9 |

**Table 2.6: ICD-9 Codes for Toxoplasmosis.** * Included for completeness of coding, but important to note that there are no reported cases of conjunctivitis secondary to *T. gondii* infection in the literature

Estimates of Annual Rates of Toxoplasmosis Making an assumption that rates of infection are identical between our study population and the general population regardless of type of insurance, which might not be accurate, and recognizing that the database represents 15% of the total population, and only those with indemnity insurance, number of cases for the

| Condition | ICD-9 Code |
|---|---|
| Communicating Hydrocephalus, | 331.3 |
| Obstructive Hydrocephalus, | 331.4 |
| Idiopathic Normal Pressure Hydrocephalus, | 331.5 |
| Congenital Hydrocephalus* | 742.3 |
| Epilepsy* | 345.x |
| Bipolar Disorder* | 296.x |
| Schizophrenia* | 295.x |
| Premature Labor, Fetal Growth Retardation* | 644.x, 764.9x |
| Alzheimer's Disease | 331 |
| Human Immunodeficiency Virus (HIV) Disease* | 42 |
| Visual Loss, Blindness, etc.* | 369.x |
| Multiple Sclerosis* | 340 |
| Benign Brain Neoplasm* | 225.0, 225.1, 225.2, 225.3, 225.4 |
| Malignant Brain Neoplasm* | 190.x |
| Anxiety* | 300.0,300.01,300.02,300.09,300.21 |
| Substance Abuse* | 304.x, 305.x |
| Thrombocytopenia* | 287.30,287.39,287.49,287.5,776.1a |
| Septicemia of the Newborn | 771.81 |
| Amyloidosis | 277.30, 277.39 |
| Memory Loss | 780.93 |
| Impulse Disorders | 312.x |
| Lupus Erythematosus, Systemic Lupus Erythematosus* | 695.4, 710.0 |
| Lymphadenopathy | 785.6 a |
| Regional Enteritis (Crohn's Disease)* | 555.x |

**Table 2.7: Comorbidities.*** Included for completeness of coding, but important to note that there are no reported cases of conjunctivitis secondary to *T. gondii* infection in the literature

entire population from 2003-2012 was calculated. Then, we subdivided by 10 years studied to give a predicted/estimated annual incidence of infection.

Comorbidity in *T. gondii* Infection Patients who have 130.x toxoplasmosis-specific codes were compared to controls matched for age, geography, and health, with fewer years in the database indicating improved health. After matching cases and controls, we constructed contingency tables for all possible toxoplasmosis-by-all-disease pairs. Using these contingency tables, we then computed the following statistics for each pair: odds ratio of disease comorbidity with toxoplasmosis codes of interest, conditional maximum likelihood estimate of disease odds ratio (with 95% confidence interval), and p-value for a null model in which two diseases occur independently of one another. We considered an association significant if it passed Benjamin-Hochberg correction [21] with a very conservative false discovery rate (FDR) threshold of 0.1%:

$$\mathrm{FDR}_r = \min_{i \geq r} \frac{p_i N}{i} \leq 0.1\%$$

where $r$ is rank of a disease ordered by increasing p-values, $p_i$ is p-value for disease with rank $i$, and $N$ is total number of diseases tested. Additional data entries in the database are in the figure legend.

## 2.4 Comorbid patterns of neuropsychiatric patients

### *2.4.1 Introduction*

Neuropsychiatric diseases are known to have significant comorbidity patterns[76, 137, 106]. Rather than being explained by biological pleiotropy, high comorbidity of Neurppsychiatric diseases are sometimes attributed to diagnostic uncertainties, genetic or environmental subtypes, or patient heterogeneity [184]. In order to tease out the casual genetic and environmental factors underlying comorbidities of neuropsychiatric diseases, we conduct a study

using large diagnostic data.

In this study, we first tried to identify the comorbidity patterns for the patients of 32 neuropsychiatric diseases. In addition, we aimed to identify directionality of the correlations. Specifically, we want to identify known genetic factors or environmental factors that are causal. Mendelian diseases, for example, can served as markers of underlying variants and are more likely to have caused the comorbidity. Similarly, infections that are significantly correlated with neuropsychiatric diseases are more likely to be causal, given the neuropsychiatric disorders do not expose patients to more infectious environments.

We hypothesized that neuropsychiatric diseases share common genetic variants and environmental factors within themselves and between them and other group of disease, which can be inferred from grouping neuropsychiatric comorbidity patterns. First we want to control for factors known to influence disease risks such as, age, gender, geographical location and patient's records in our database. After these factors are controlled for, we can then assess the comorbid patterns and possibly have a better understanding of the underlying etiology of those diseases. For each pair of neuropsychiatric and comorbid disease, the hypotheses tested are the pair could be independent, positively correlated or negatively correlated.

For each group of neuropsychiatric disease patients, a group of covariates matched controls are sampled. The covariates included for matching are age, gender, county and number of years in our databased. Then Fisher's exact test is used to calculate the odds ratio of each pair of neuropsychiatric disease and comorbid disease. We used three matching schemes to test the robustness of the correlations. Perfect matching includes only patients with exact same covariates therefore are the fastest, most stringent and least powered method. Mahalanobis distance scheme finds the closest matched controls using covariates covariance matrix. The propensity score matching matches individuals based on their probability of being diagnosed with the neuropsychiatric disease of interest. Propensity score matching analysis generates the most number of significant results and its results are reported here.

### *2.4.2 Results and Discussion*

Here we report the significant patterns of neuropsychiatric patients' comorbidity in major biological systems.

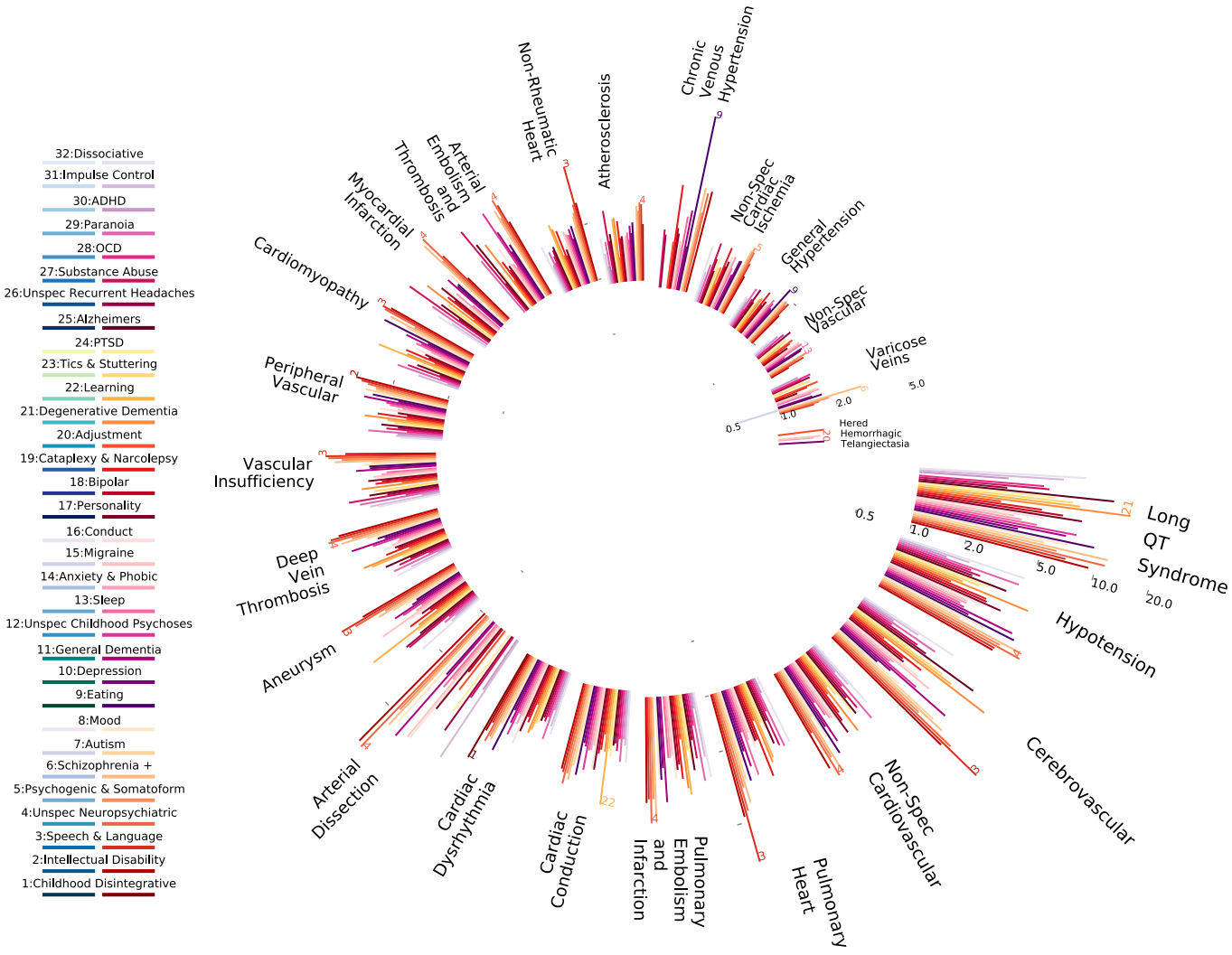


**Figure 2.11: Cardiovascular diseases odds ratios of Neuropsychiatric patients versus Control.**

**Cardiovascular Diseases** Comparing to control patients, neuropsychiatric patients in general have higher odds ratios for cardiovascular diseases (as high as ten-fold increase). Interestingly, Long QT syndrome odds are increased in several neuropsychiatric diseases patients. However, there exists a significant negative correlation between Varicose Veins and

Autism.

Positive correlations highlights:

- Degenerative Dementia and Alzheimers → Long QT Syndrome possibly mediated by drug side effects [321]
- Eating disorder → Chronic Venous Hypertension
- Speech & Language ← Cerebrovascular Diseases
- Speech & Language ⟷ Pulmonary Heart Disease. This association may be caused by a common factor, dysphagia [271]
- Schizophrenia, Unspecific Neuropsychiatric Disorder ⟷ Hypotension, Pulmonary Embolism and Infarction [76, 168, 279]

Negative correlations highlights:

- Autism ⟷ Varicose Veins

**Central Nervous system** Strong positive correlations between diagnosis of neuropsychiatric diseases and central nervous system diseases. We did not observe any negative correlations. Degenerative Dementia and Intellectual disability are the two neuropsychiatric diseases that are highly positive correlated with CNS diseases.

Positive correlations highlights(over 150 fold increase):

- General dementia ← Specified Childhood Cerebral Degeneration
- Unspecified Childhood Psychosis ⟷ Specified Childhood Cerebral Degeneration
- Intellectual disability← Specified Childhood Cerebral Degeneration, Unspecified Childhood Cerebral Degeneration

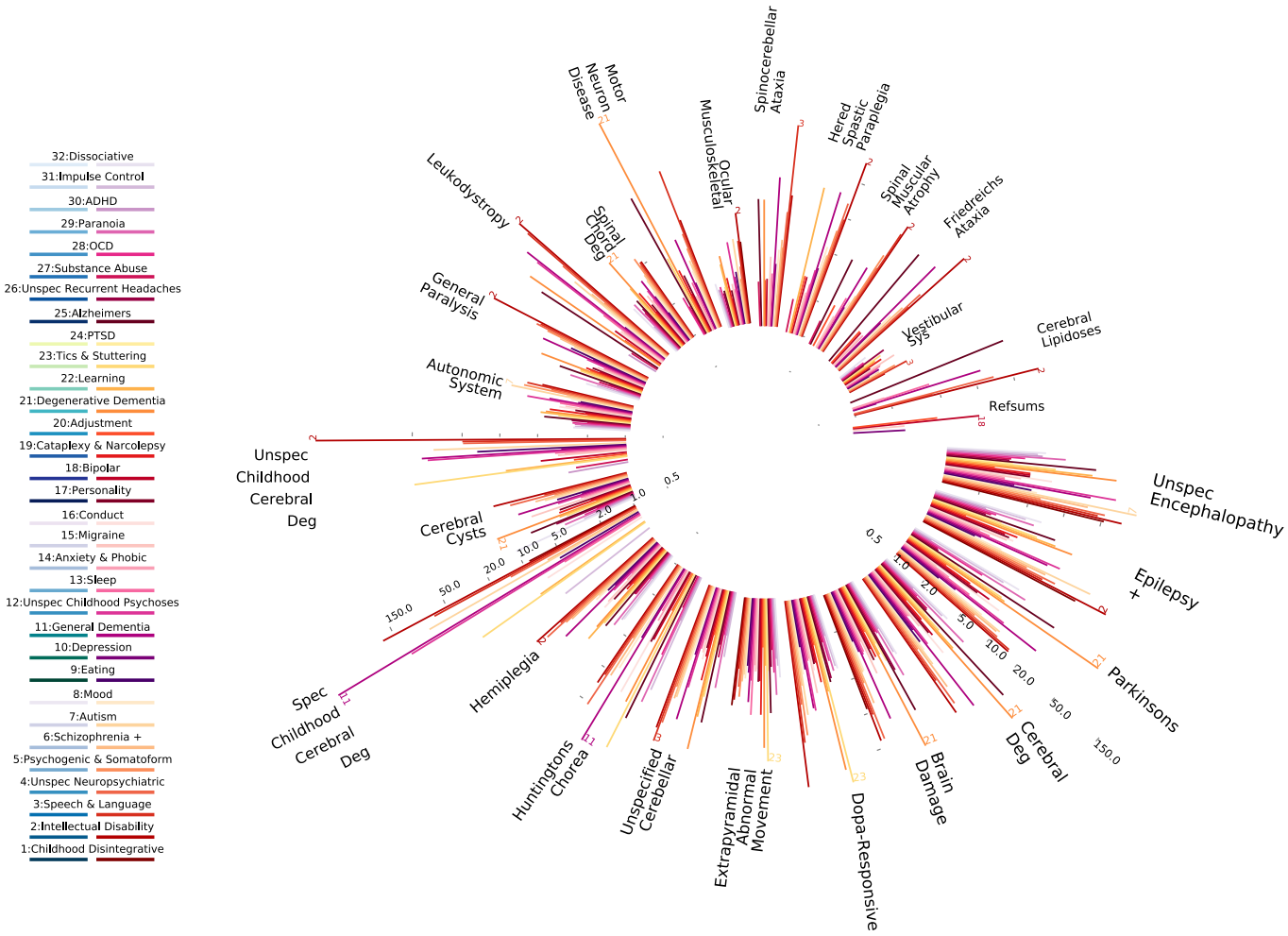


**Figure 2.12: Central Nervous system diseases odds ratios of Neuropsychiatric patients versus Control.**

- Intellectual disability⟷ Epilepsy, Cerebral Lipidoses, Ataxia, Spinal Muscular Atropy, Leukodystropy, General Paralysis. These associations are likely caused by epistatic Mendelian factors [110, 313].

- Degenerative Dementia ⟷ Parkinsons, General Cerebral Degeneration, Brain damage, Cerebral Cysts, Motor neuron Disease

**Developmental system** Both strong postive correlaitons and strong negative correlation were presented in the developmental system diseases of neuropsychiatric patients. For example, Migraine, Bipolar, Substance abuse patients have a significant lower change(as low as

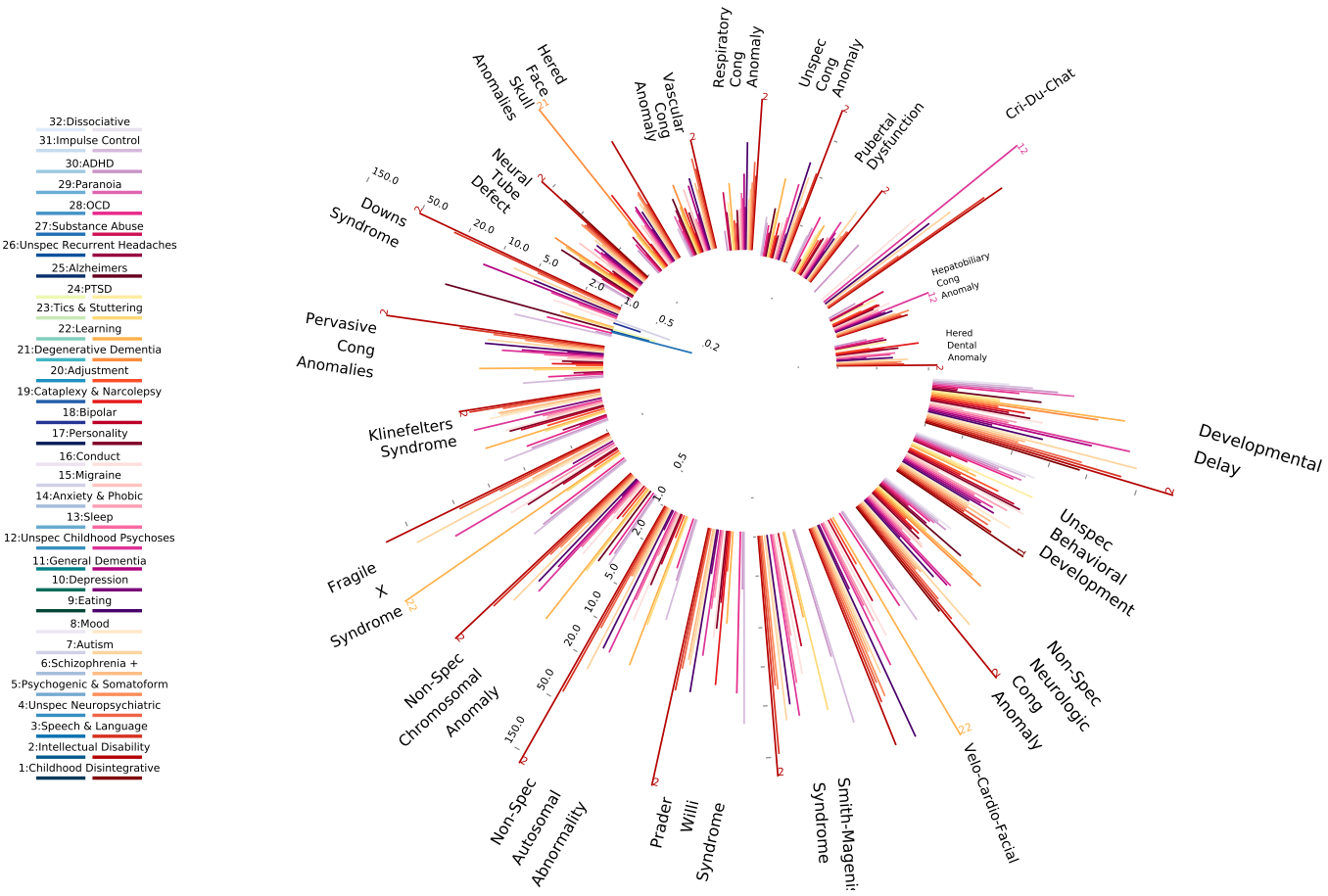


**Figure 2.13: Developmental diseases odds ratios of Neuropsychiatric patients versus Control.**

0.23) to be a Downs syndrome patient. On the other hand, several developmental diseases are positive correlated with intellectual disability, possibly mediated by Downs syndrome and other chromosomal disorders.

Positive correlations highlights(over 150 fold):

- Non-specific Chromosonal/autosomal abnormality → intellectual disability Fragile X syndrome, Velo-Cardio-Facial syndrome → Learning disorders

Negative correlations highlights:

- Downs syndrome, Balanced Translocation ⟷ Substance abuse(0.23)
- Downs syndrome → Migraine and Bipolar

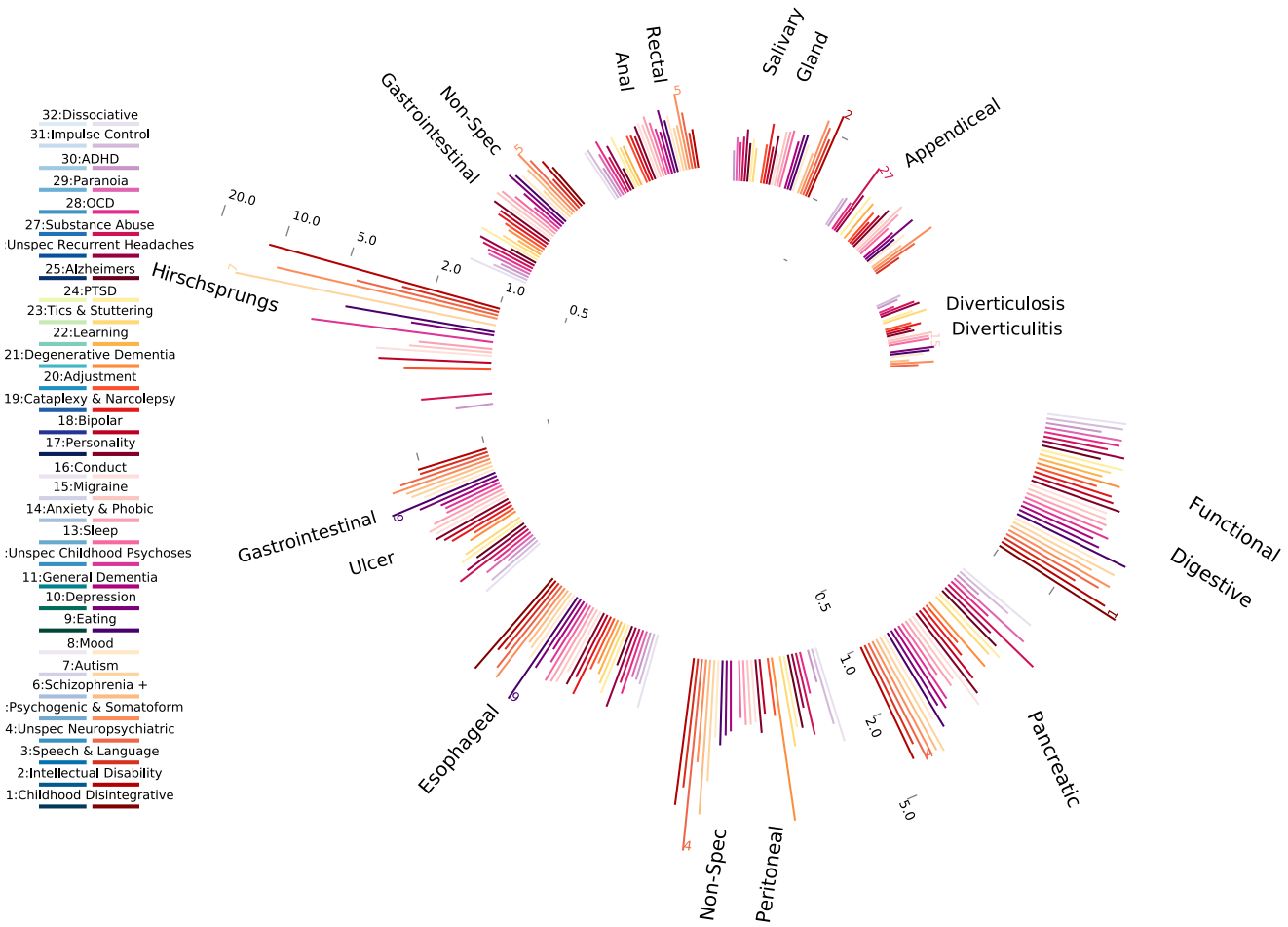


**Figure 2.14: Digestive diseases odds ratios of Neuropsychiatric patients versus Control.**

**Digestive system** Neuropsychiatric patients have a relatively low positive correlations with digestive system diseases

Positive correlations highlights:

- Autism, Intellectual Disability ⟷ Hirschsprungs. These associations could be mediated by Down syndrome [84]

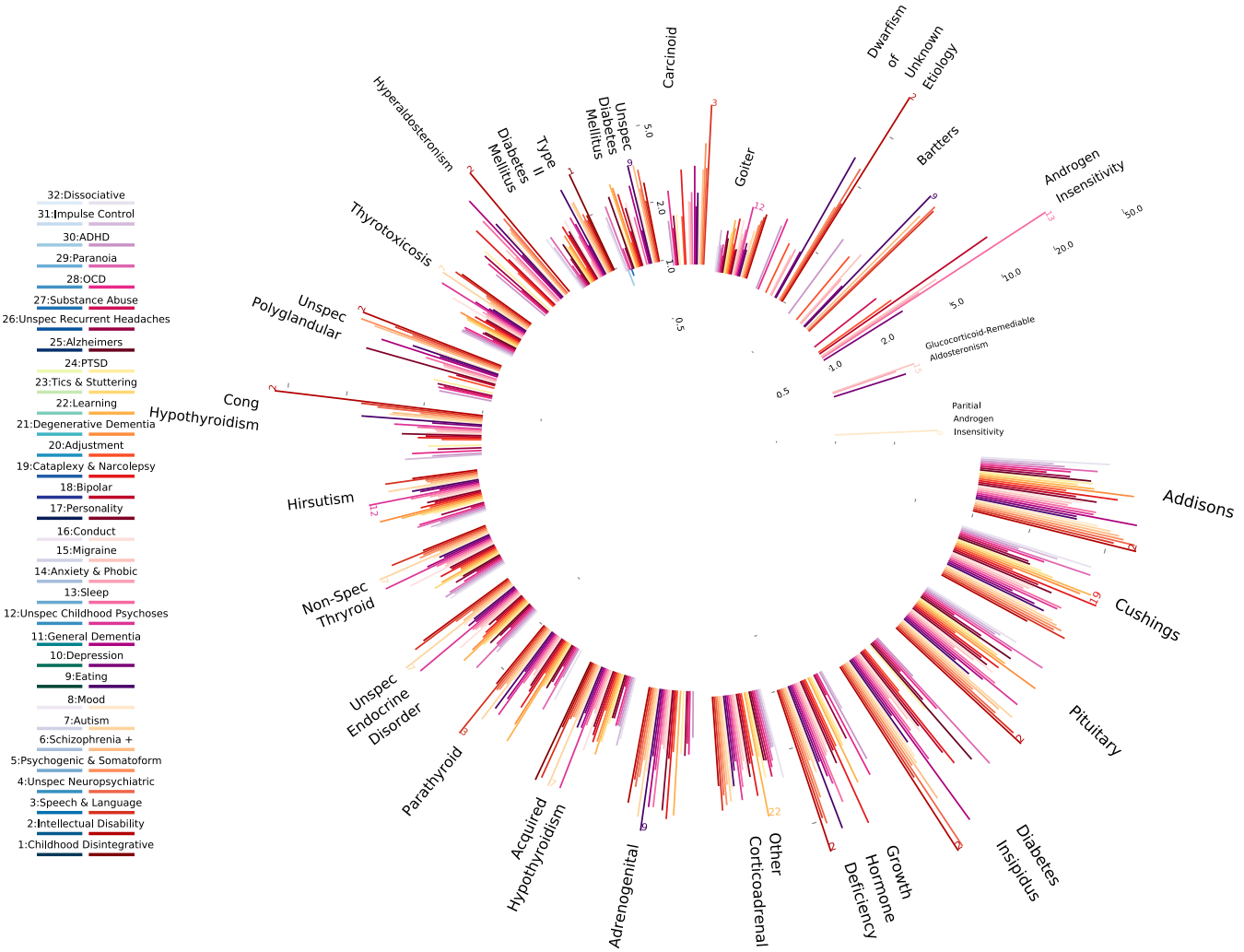


**Figure 2.15: Endocrine diseases odds ratios of Neuropsychiatric patients versus Control.**

**Endocrine system** In general, Neuropsychiatric diseases have very moderate positive correlations with endocrine system diseases with the exception of Androgen Insensitivity Bartters syndrome and Dwarfism.

Positive correlations highlights:

- Androgen Insensitivity → Sleep disorder and Bipolar disorder, possibly caused by epistatic effects of Mendelian genetic factors.

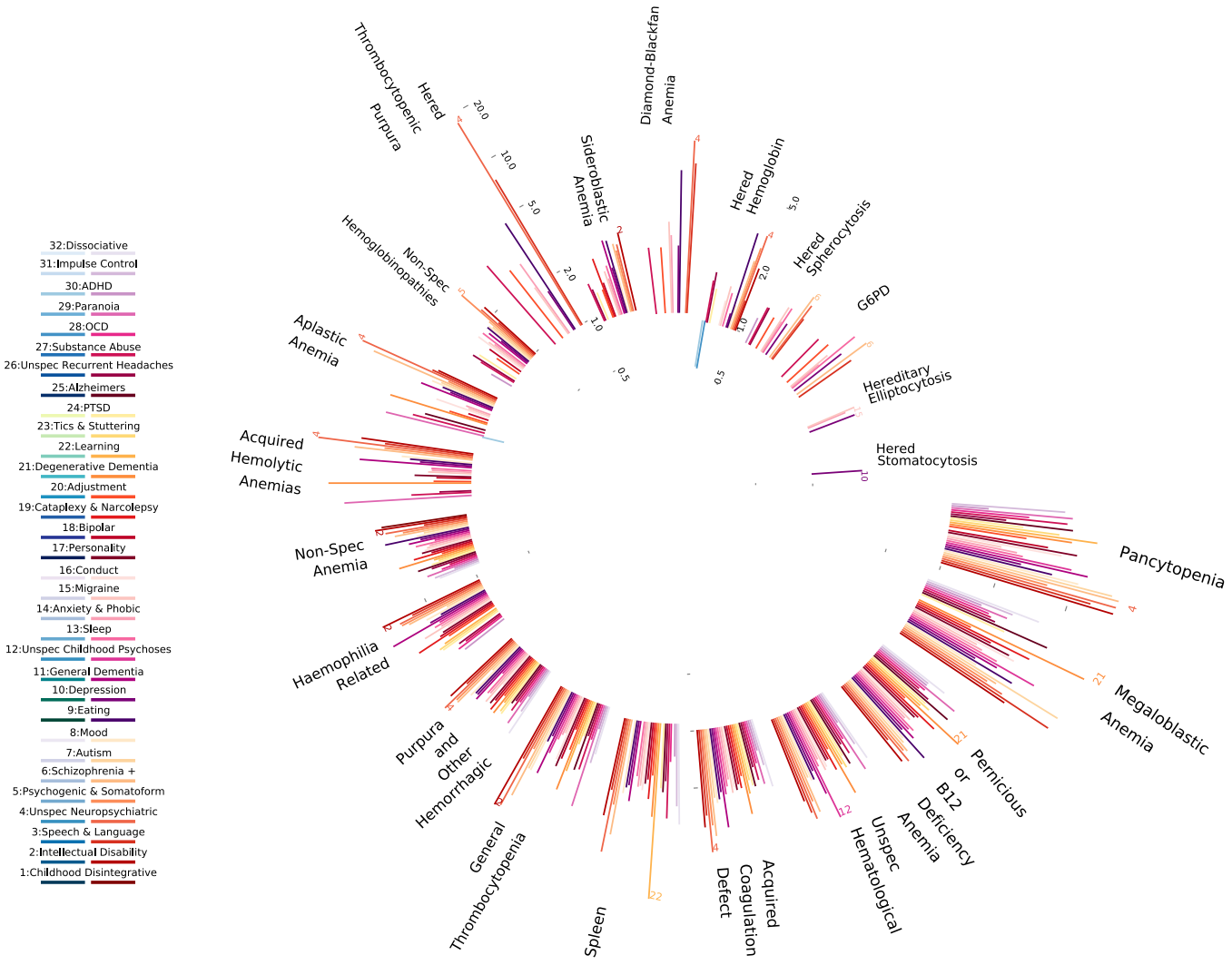


**Figure 2.16: Hematologic diseases odds ratios of Neuropsychiatric patients versus Control.**

**Hematologic system** For Hematologic diseases, there are low positive correlations with a few negative correlations The main positive correlations is from Unspecified Neuropsychiatric disorder, it may have something to do with code 306.4 Gastrointestinal malfunction arising from mental factors.

Positive correlations highlights:

- Unspecific Neuropsychiatric disorder ← Hereditary thrombotic thrombocytopenic purpura (TTP), Pancytopenia, Anemias, and Acquired Coagulation Defect

- Degenerative Dementia ← Anemias

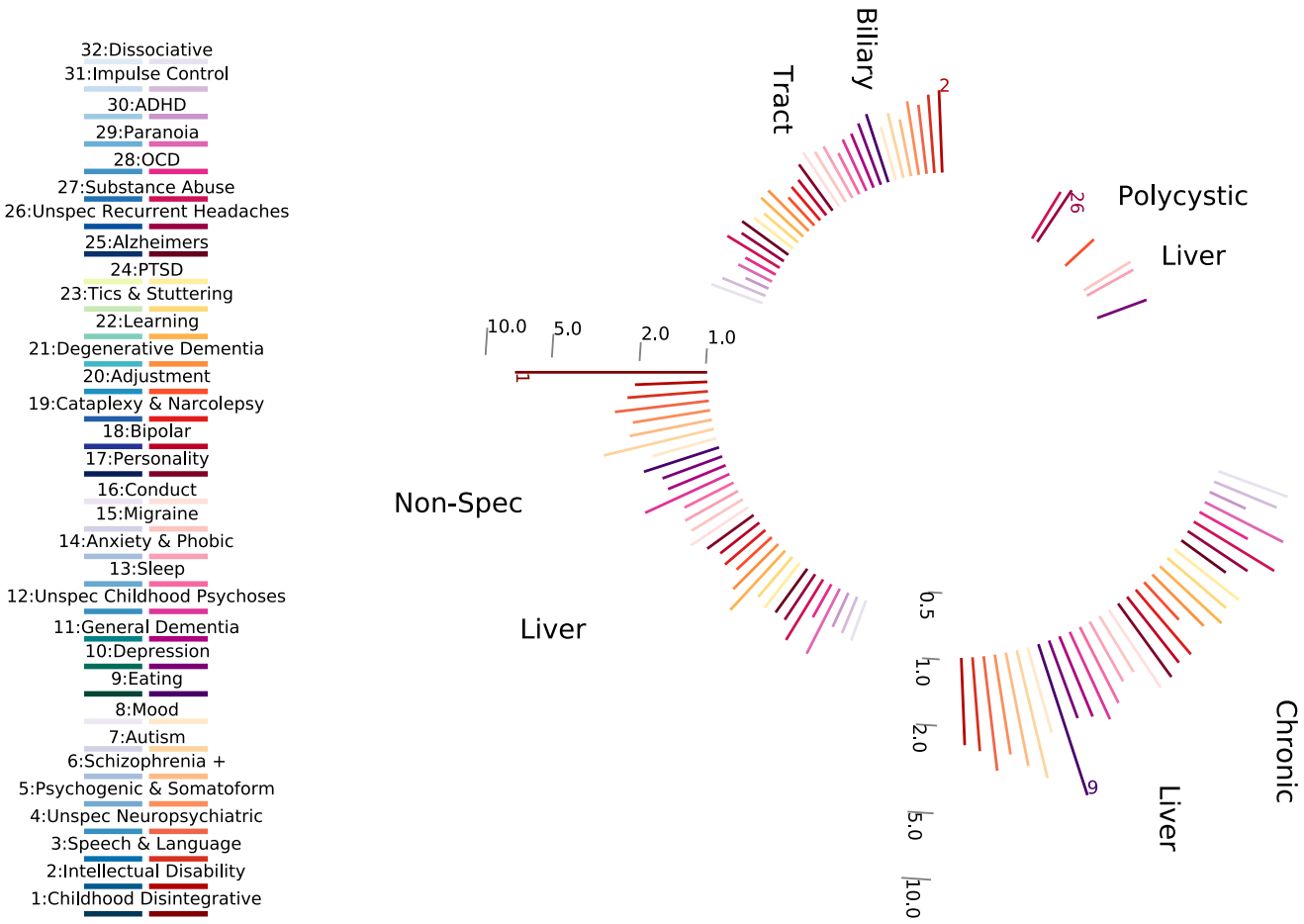


**Figure 2.17: Hepatic diseases odds ratios of Neuropsychiatric patients versus Control.**

**Hepatic system** Although smaller number of Hepatic diseases are correlated with neuropsychiatric disorders, several interesting cases of the correlations can be found.

Positive correlations highlights:

- Childhood Disintegrative disorder ⟷ Non-specific Liver disease
- Eating disorder → Chronic Liver Disease [191]

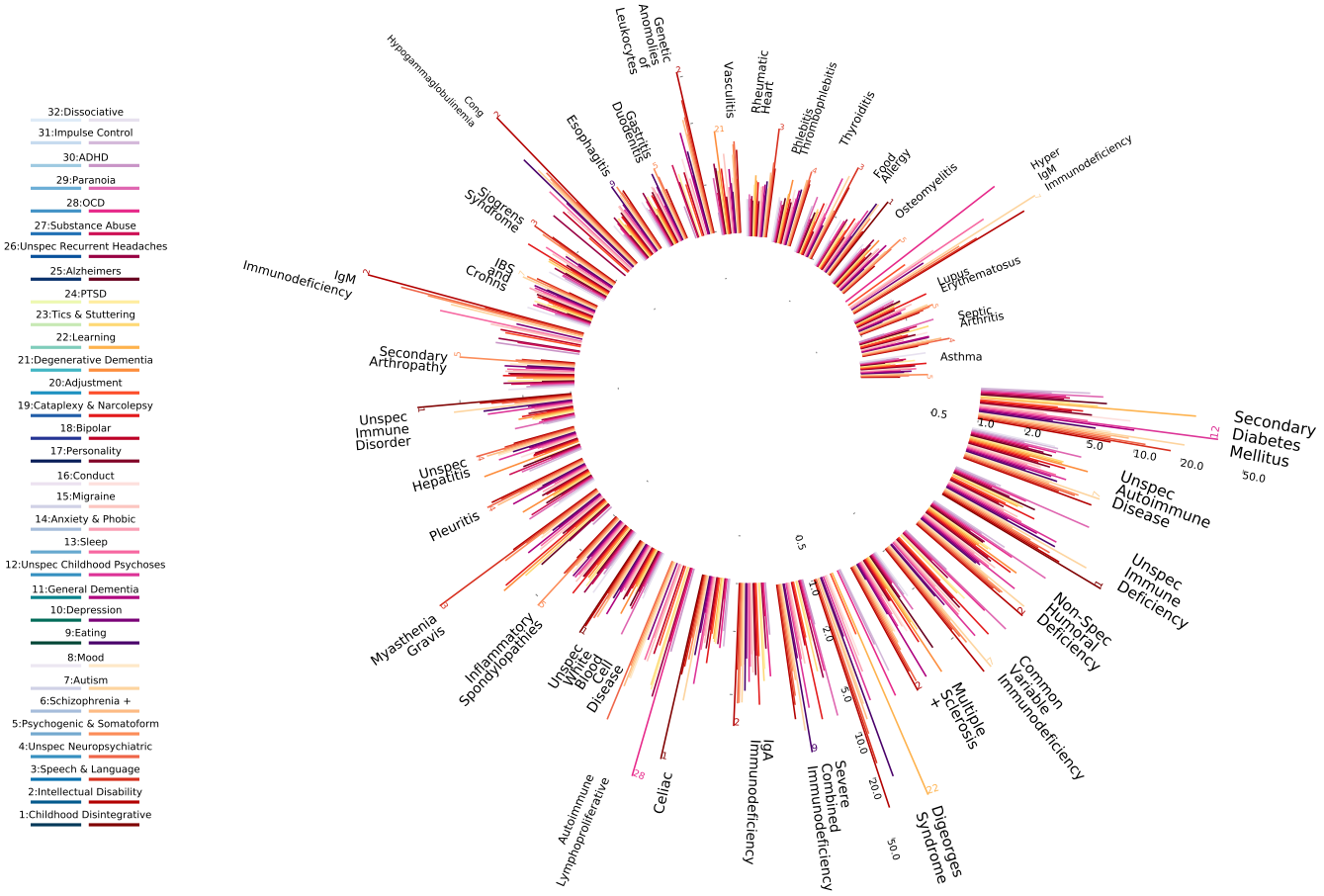


**Figure 2.18: Immunological diseases odds ratios of Neuropsychiatric patients versus Control.**

**Immunological system** In general, there are relative strong positive correlations between Immunological diseases and neuropsychiatric diseases. while no negative correlation was observed. Most Immunological diseases are correlated with Intellectual Disability and Speech and Language Disorders

Positive correlations highlights

- Unspecific Childhood Psychoses⟷Secondary Diabetes Mellitus

- Childhood Disintegrative Disorder, Learning disorder ← Digeorges Syndrome [58]
- OCD ← Autoimmune Lymphoproliferative Syndrome(Mendelian)
- Speech and Language Disorder ⟷ Myasthenia Gravis
- Intellectual Disability ⟷ IgM Immunodeficiency, most likely mediated by Down syndrome
- Autism, Childhood Disintegrative Disorder, OCD ⟷ Hyper IgM Immunodeficiency (X-linked recessive)
- Speech and Language Disorder ⟷ Autoimmune Polyglandular Disorder, Cyclic Neutropenia
- Unspecific Neuropsychiatric disorder, Psychogenic & Somatoform disorder ← Familial Mediterranean Fever(Mendelian)

**Infectious Diseases** Not surprisingly Prion disease have several significant correlations with neuro-degenerative diseases (Alzheimer disease, dementias). However, in general the positive odds ratios of infectious diseases in neuropsychiatric patients are relatively small. Degenerative Dementia and Unspecific Neuropsychiatric disorder have several positive correlations with other infectious diseases.
Positive correlations highlights

- Alzheimer's disease, General Dementia, Unspecified Neuropsychiatric disorders ← Prion
- Degenerative Dementia ← Encephalitis, E. coli Infection, Meningitis, HIV
- Unspecific Neuropsychiatric disorder ← Staph Infection, CNS Infection, Septicemia, Viral Hep C, Connetive Tissue Infection, Carditis, Peritonitis and Retroperitoneal Infection

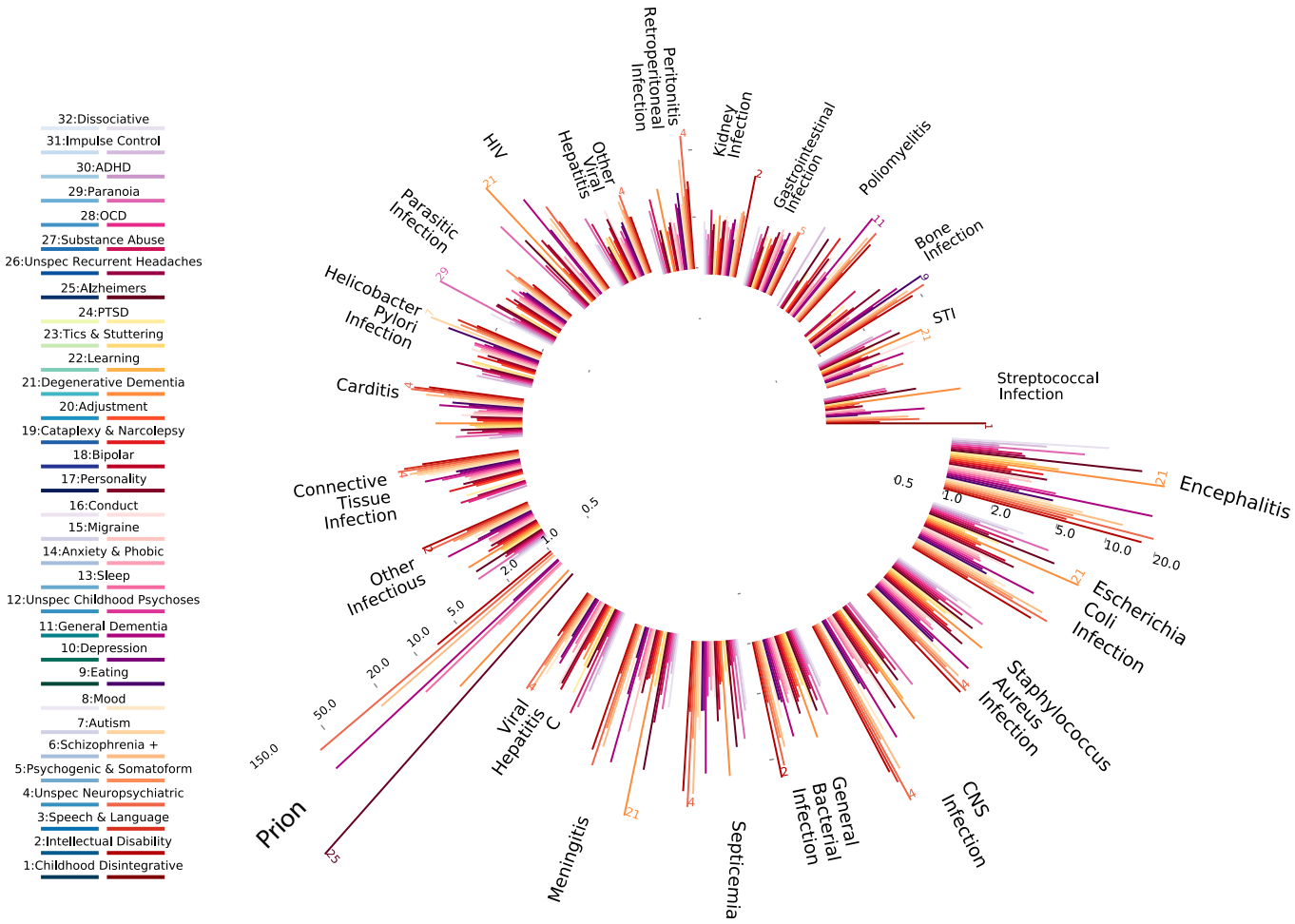


**Figure 2.19: Infectious diseases odds ratios of Neuropsychiatric patients versus Control.**

**Metabolic system** Some of the positive correlations between Metabolic diseases and neuropsychiatric diseases are as high as 50 fold, most of them are with Intellectual disability, likely mediated by the epstatic effects of Down syndrome

Positive correlations highlights

- Intellectual Disability $\longleftrightarrow$ Urea Cycle Metabolism, Mitochondrial Metabolism, Lysosomal Storage, Primary Carnitine Deficiency, Fatty Acid Oxidation, Iron Metabolism, Branched Chain AminoA Metabolism

- Childhood Disintegrative Disorder $\longleftrightarrow$ Non-Specific Metabolic Disorder

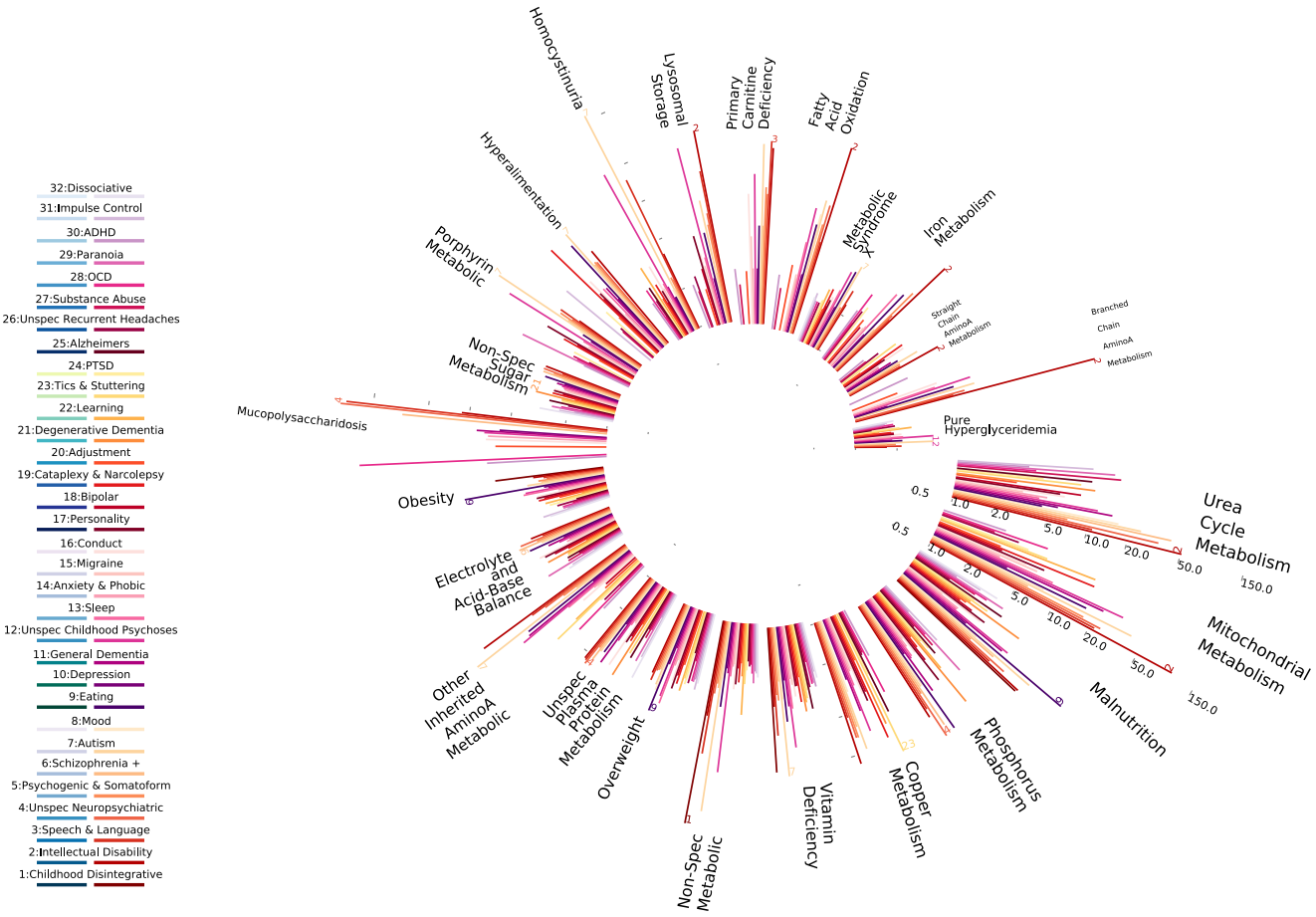


**Figure 2.20: Metabolic diseases odds ratios of Neuropsychiatric patients versus Control.**

- Eating Disorder → Malnutrition, Localized Adiposity
- Speech & Language, Unspec Neuropsychiatric ⟷ Mucopolysaccharidosis
- Autism ⟷ Homocystinuria, Vitamin Deficiency
- Speech Language ⟷ Peroxisomal Disorders, Alpha-1-Antitrypsin Deficiency
- Intellectual Disability ⟷ Galactosemia
- Learning Disorder ⟷ Glycogenosis
- Intellectual Disability, Speech & Language ← Purine Pyrimidine Metabolic

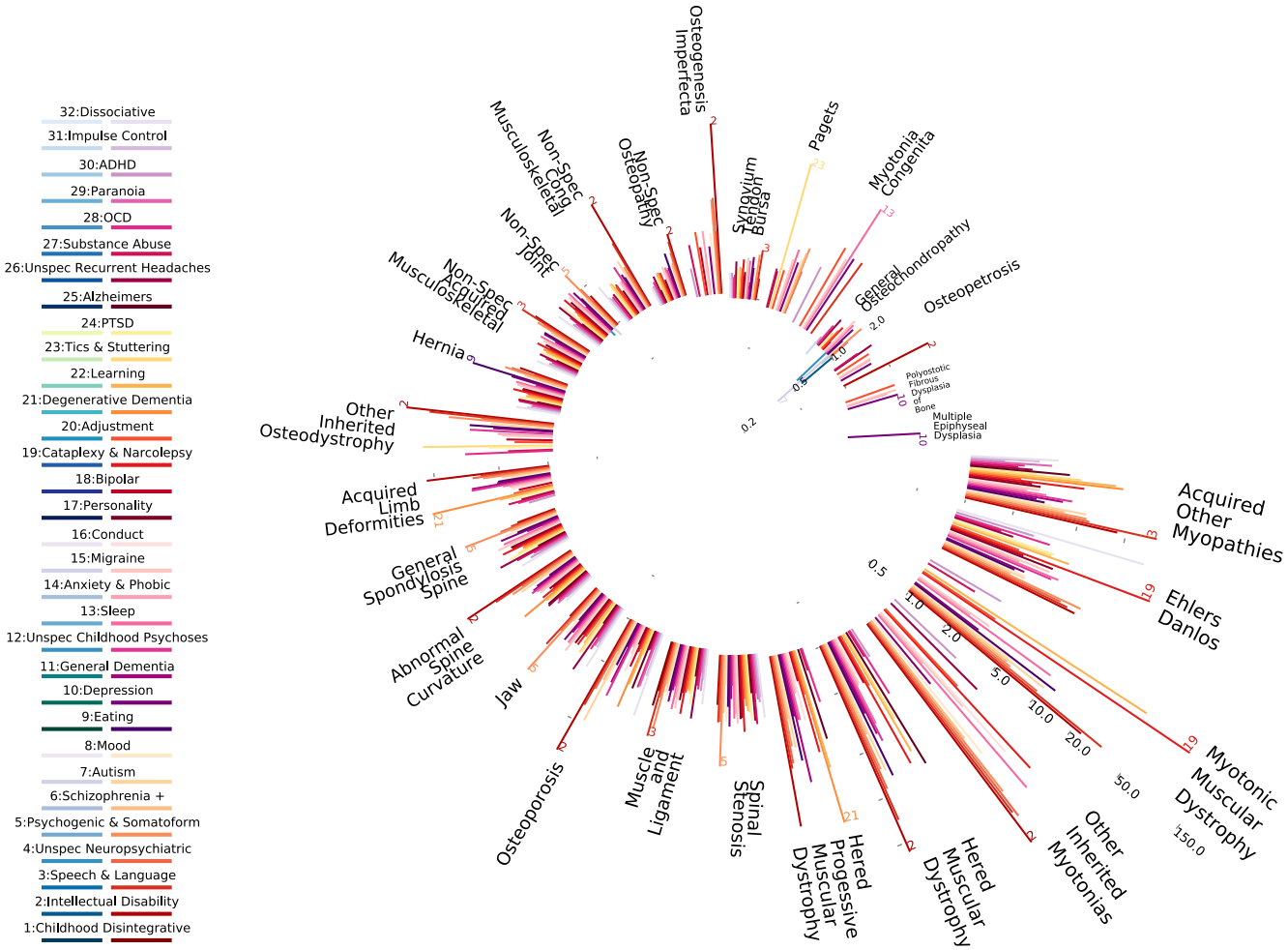


**Figure 2.21: Musculoskeletal diseases odds ratios of Neuropsychiatric patients versus Control.**

**Musculoskeletal** Mostly low odds ratios increase for neuropsychiatric patients, with a few exceptions, Myotonic Muscular Dystrophy and Other Inherited Myotonias are highly correlated with Cataplexy and Narcolepsy.

Positive correlations highlights

- Cataplexy and Narcolepsy ← Myotonic Muscular Dystrophy, Ehlers Danlos

**Neoplastic Process** Neoplastic diseases have the largest number of inverse comobidity, meaning two diseases are negative correlated. Some of these negative correlations are well documented in literature while several are newly discovered.

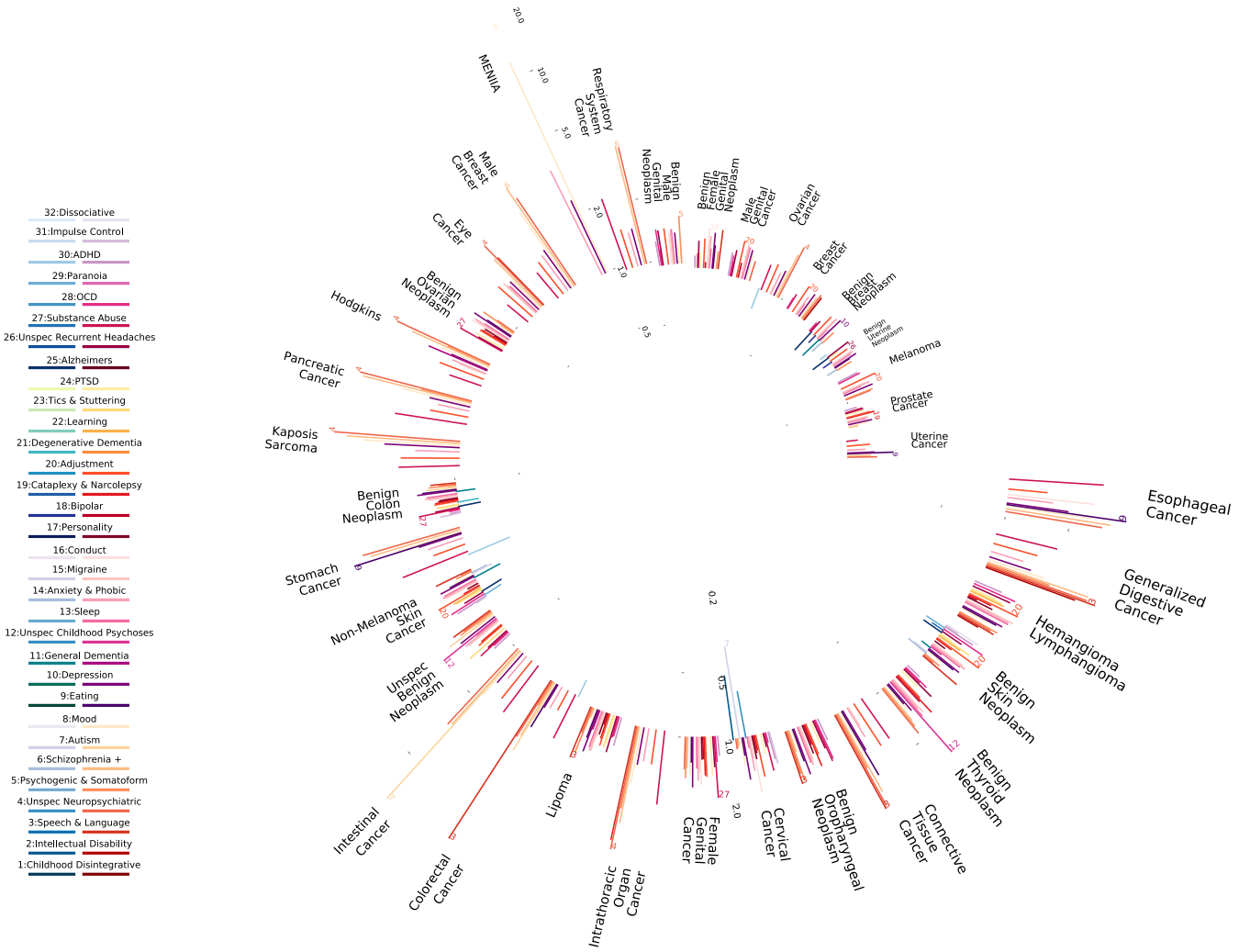


**Figure 2.22: Neoplastic Process diseases odds ratios of Neuropsychiatric patients versus Control.**

Positive correlations highlights

- Intellectual Disability $\longleftrightarrow$ Tuberous Sclerosis
- Autism $\longleftrightarrow$ Ruvalcaba-Myhre, Intestinal Cancer
- Mood disorder $\longleftrightarrow$ MENIIA
- Speech and Language $\longleftrightarrow$ Colorectal Cancer

Negative Highlights

- Speech and Language, Autism ⟷ Cervical Cancer
- Alzheimer, General Dementia, Degenerative Dementia ⟷ Benign Colon Neoplasm
- Alzheimer, Schizophrenia ⟷ Benign Breast Neoplasm, Benign Uterine Neoplasm, Non-Melanoma Skin Cancer, Benign Skin Neoplasm [135]
- ADHD ⟷ Stomach Cancer, Colorectal Cancer, Lung Cancer, Secondary Malignant Neoplasm,Ovarian Cancer

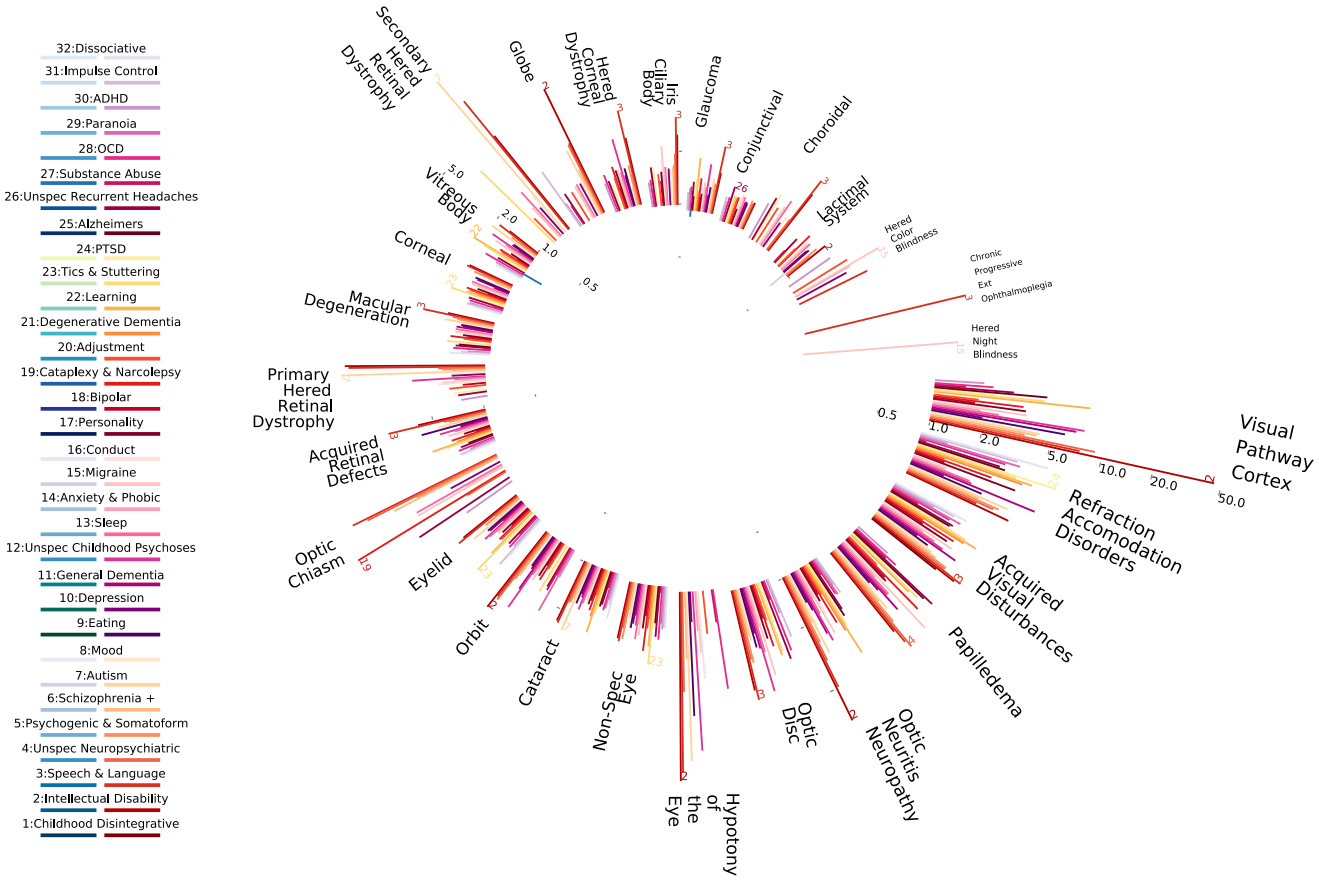


**Figure 2.23: Ophthalmological diseases odds ratios of Neuropsychiatric patients versus Control.**

**Ophthalmological** Among the relatively low positive correlations between Ophthalmological diseases and neuropsychiatric diseases are a few high positive correlations, with over

50 fold increase of odds ratio for Visual Pathway Cortex Disorder in Intellectual Disability patients

Positive correlations highlights:

- Intellectual Disability ⟷ Visual Pathway Cortex Disorder, Hypotony of the Eye, Optic Neuritis Neuropathy
- Autism ← Primary Hereditary Retinal Dystrophy, Secondary Hereditary Retinal Dystrophy
- Cataplexy and Narcolepsy ⟷ Optic Chiasm Disorder

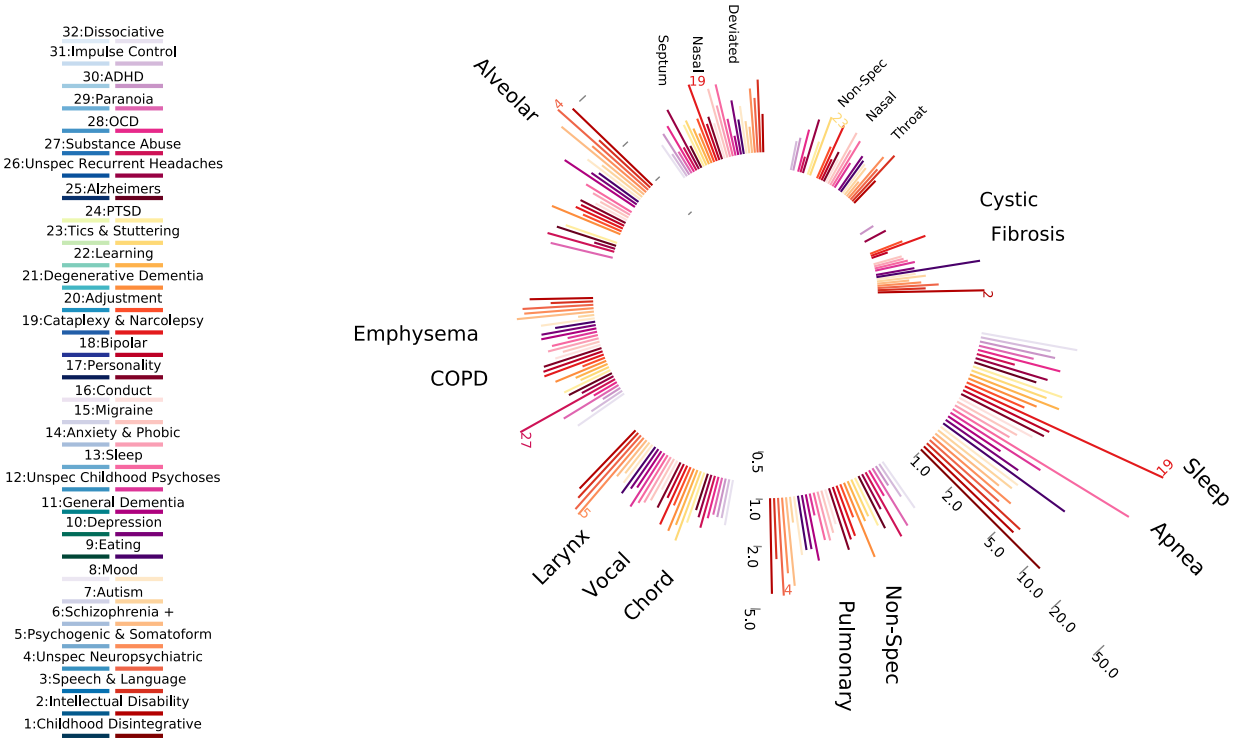


**Figure 2.24: Respiratory diseases odds ratios of Neuropsychiatric patients versus Control.**

**Respiratory System** Previous literature has found several correlations between Cystic Fibrosis and neuropsychiatric diseases. Our analysis confirmed these findings while uncover several novel positive correlations between respiratory disorders and neuropsychiatric disorders. Positive correlations highlights:

- Cataplexy & Narcolepsy, Childhood Disintegrative Disorder,Eating Disorder $\longleftrightarrow$ Sleep Apnea
- Substance Abuse $\rightarrow$ Emphysema COPD
- Schizophrenia, Unspec Neuropsychiatric Disorder,Intellectual Disability $\longleftrightarrow$ Alveolar Disease
- Eating Disorder $\leftarrow$ Cystic Fibrosis [231, 233]
- Intellectual Disability $\leftarrow$ Cystic Fibrosis, likely caused by malnutrition [155]

**Urinary System** Several moderate positive correlations (more than 10 fold) stand out among relative few and low correlations between urinary system diseases and neuropsychiatric diseases.
Positive correlations highlights:

- Speech and Language $\leftarrow$ Renal Glycosuria(Mendelian)
- Psychogeneic and Somatoform Disorder $\leftarrow$ Medullary Cystic Kidney(Mendelian)
- Schizophrenia, Unspecific Neuropsychiatric Disorder $\longleftrightarrow$ Acute Renal Failure, Chronic Kidney Disease [290, 129]
- Eating Disorder $\longleftrightarrow$ Polycystic Kidney(Mendelian)

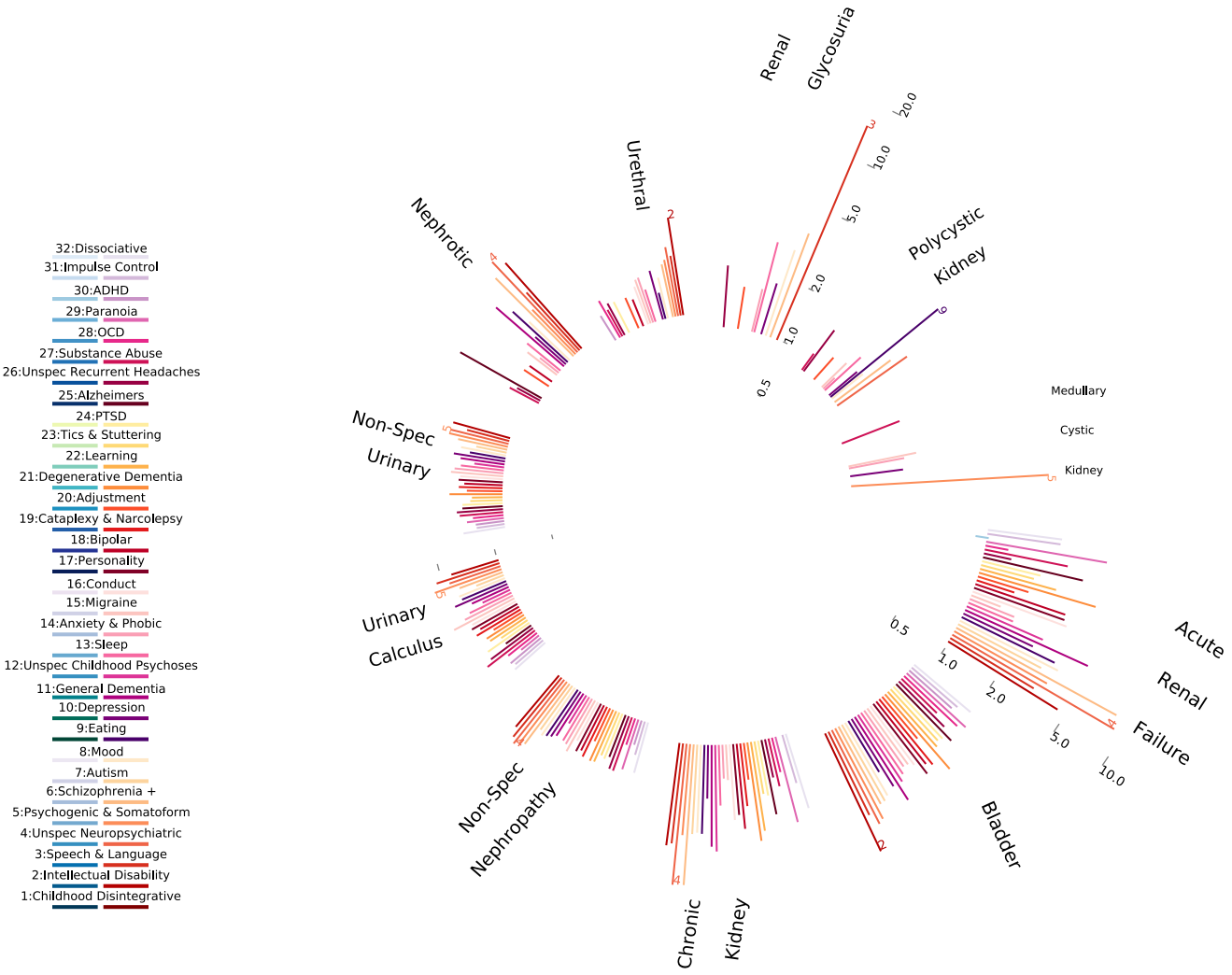


**Figure 2.25: Urinary diseases odds ratios of Neuropsychiatric patients versus Control.**

### *2.4.3 Methods*

## Data Sources

We used the Truven Health Analytics MarketScan® Commercial Claims and Encounters Database. It spans the years 2003 to 2013. It consists of approved commercial health insurance claims for between 17.5 and 45.2 million people annually, with linkage across years, yielding a total of approximately 115 million patient records. This national database contains information contributed by well over 100 insurance carriers and large self-insuring companies. With respect to this dataset, we scanned approximately 4.6 billion inpatient and outpatient service claims and identified almost 6 billion diagnosis codes. After removing duplicates, almost 1.3 billion diagnosis codes were found to be associated with over 99.1 million individuals, yielding approximately 12.89 unique diagnostic codes per individual.

## Case Control Matching

**Rubin Causal Model** In randomized experiments individuals are randomly assigned to treatment or control group to ensure balance of the covariates between the two groups. In nonexperimental studies, unadjusted results would suffer from confounding by indication that would be difficult to address since the number of variables that might differ between cases and controls is large. To address these concerns, this study employed a matched cohort design based on the Rubin causal model. A key assumption in applying this model to nonexperimental studies is that of a strongly ignorable treatment; treatment assignment ($T$) is independent of the potential outcomes($Y(0)$, $Y(1)$) given the covariates($X$): $T \perp (Y(0), Y(1))|X$. In general, matching typically proceeds by including an unexposed individual with the same covariate level (or patterns of covariates) for each exposed individual with a given covariate level.

To formalize, using notation similar to that in [241], we consider two populations, $P_t$ and $P_c$, where the subsubscript $t$ refers to a group exposed to the treatment and $c$ refers to a

group exposed to the control. Covariate data on $p$ pre-treatment covariates is available on random samples of sizes $N_t$ and $N_c$ from $P_t$ and $P_c$. The means and variance covariance matrix of the $p$ covariates in group $i$ are given by $\mu_i$ and $\Sigma_i$ , respectively ($i = t$ , $c$). For individual $j$ , the $p$ covariates are denoted by $\boldsymbol{x_j}$ , treatment assignment by $T_j$ ($T_j = 0$ or 1), and the observed outcome by $Y_j$ . Without loss of generality, we assume $N_t < N_c$.

**Covariates and Variables** The clinical record database was first parsed, then each patient is characterized by five covariates for the case-control matching and considered all diagnostic variables for comparison.

To control for age related effects, we used age reported in the database as the first covariate. This usually indicates the year a patient enters the database. Gender is used as the second covariate. To control for geographic confounding factors, we used the county a patient first reported in the database as a covariate.

We manually curated a dictionary of diseases and their corresponding ICD9 codes. Each disease variable is composed of a set of ICD9 terms. For example, the variable "Autism" is composed of three different ICD9 concepts, "299.0: Infantile autism, residual state", " 299.00: Autistic disorder" and "299.01: Infantile autism, current or active state". Each disease was categorized based on its general affected system, applicable gender, and inheritance. In our matching scheme, we selected all neuropsychiatric diseases as "treatment" variables. For each neuropsychiatric disease, all other diagnosed diseases in a patient were considered "outcome" variables. The list of diseases and concepts defining the diseases used in this study was manually curated.

**Distance Calculations** To measure the similarity between two indiviudals. There are three primary ways to define the distance $D_{ij}$between individuas $i$ and $j$ for matching.

1. Exact:

$$D_{ij} = \begin{cases} 0, & \text{if } \boldsymbol{x_i} = \boldsymbol{x_j}, \\ \infty, & \text{if } \boldsymbol{x_i} \neq \boldsymbol{x_j}. \end{cases}$$

One of the strictest ways of eliminating confounding effects is to find controls that have the exact same covariates as cases. To construct such control group, we selected patients free of the neuropsychiatric disease and had perfect matches for all covariates. We matched cases and controls using three different exact match schemes: one to one , one to five and one to all available nearest neighbor matching.

2. Mahalanobis:

$$D_{ij} = (\boldsymbol{x_i} - \boldsymbol{x_j})' \boldsymbol{\Sigma^{-1}} (\boldsymbol{x_i} - \boldsymbol{x_j}).$$

Here we chose to focus on the average effect of the treatment on the treated (ATT), where $\Sigma$ is the variance covariance matrix of $X$ in the full control group. We matched cases and controls using the one to five nearest neighbor matching schemes.

3. Propensity score:

$$D_{ij} = |e_i - e_j|,$$

where $e_k$ is the propensity score for individual $k$, defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates. The probability of being in the treatment:

$$e_i = Pr(T_i = 1 | \boldsymbol{x_i})$$

To model the propensity score we use logistic regression model, we let

$$ln \frac{Pr(T_i = 1 | \boldsymbol{x_i})}{1 - Pr(T_i = 1 | \boldsymbol{x_i})} = \beta^T \boldsymbol{x_i} = \beta_0 + \beta_1 \times C_i + \beta_2 \times G_i + \beta_3 \times A_i.$$

while the parameter $\beta = (\beta_0, ..., \beta_p)$ is a $(p + 1) \times 1$ vector of regression coefficients. And $C_i, G_i$, and $A_i$ are the county, gender, and age covariates.

**Matches Diagnostics** One of the most common numerical balance diagnostics is the "standardized bias", difference in means of each covariate, divided by the standard deviation

in the full treated group:

$$d = \frac{\mid \overline{X}_t - \overline{X}_c \mid}{\sigma_t}.$$

The second diagnostics we used is standardized difference defined as:

$$d = \frac{\mid \overline{X}_t - \overline{X}_c \mid}{\sqrt{\frac{s_t^2 + s_c^2}{2}}}.$$

The third diagnostics is the ratio of the variances for each covariate. Based on [242] the absolute standardized differences of means should be less than 0.25 and the variance ratios should be between 0.5 and 2. In addition, a standard difference that is less than 0.1 has been taken to indicate a negligible difference in the mean or prevalence of a covariate between treatment groups [207].

## Statistical Tests

After matching the case and control groups, we constructed contingency tables for all possible neuropsychiatric-by-all disease pairs. Using these contingency tables, we then computed the following statistics for each pair: the odds ratio of the disease co-morbidity with neuropsychiatric diseases of interest, the conditional maximum likelihood estimate of the disease co-morbidity odds ratio (with 95% confidence interval), and the p-value for a null model in which the two diseases occur independently of one another. We considered an association significant if it passed the Benjamini—-Hochberg correction [22, 23]with a very conservative FDR threshold of 0.1%:

$$FDR_r = \min_{i \geq r} \frac{p_i N}{i} \leq 0.1\%$$

where $r$ is the rank of a disease ordered by increasing p-values, $p_i$ is the p-value for disease with rank $i$, and $N$ is the total number of diseases tested.

## 2.5 Pre-existing type 2 immune activation protects against the development of sepsis

### *2.5.1 Main*

Sepsis is a clinical syndrome of life-threatening organ dysfunction caused by a dysregulated host immune response to infection[261]. This response typically involves overwhelming type 1/Th1 or type 17/Th17 immune-mediated inflammation to promote pathogen clearance[222]. However, these inflammatory responses also mediate the organ dysfunction characteristic of sepsis, causing significant tissue damage that the body must repair[125, 91]. Little is known about counter-regulatory reparative responses during sepsis, although an overly robust anti-inflammatory response may contribute to sepsis-associated death[124]. Given the importance of immune responses in acute sepsis pathophysiology, we speculated that the immune responses associated with pre-existing comorbidities would impact the development of sepsis during acute infection. Using the Truven MarketScan private insurance claims database of over 150M people, we identified 73,587 patients with sepsis based on International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes for sepsis (995.91) or severe sepsis (995.92). We matched 5:1 non-septic patients to septic patients by age, gender, location and number of comorbid diagnoses, and then determined the prevalence of various immune-mediated diseases in both groups. Odds ratios for having specific comorbid immune diseases were calculated, and significance was determined after accounting for multiple testing comparisons. Interestingly, septic patients were more likely to have a number of immune-mediated diseases, including vasculitis, ulcerative colitis, multiple sclerosis, type 1 diabetes, and (among females) lupus erythematosus (Table 2.8). However, to our surprise, diseases associated with an overactive type 2/Th2 immune response, such as asthma, allergic rhinitis, atopic dermatitis, and food allergy, were markedly and significantly underrepresented among septic patients. This remarkable discovery suggested to us that these diseases protect patients from becoming septic. To determine the biological plau-

sibility of this hypothesis, we first induced pulmonary type 2 inflammation in two mouse models of allergic asthma. Innate type 2 immune responses were induced by 3 days of intratracheal IL-33 administration, which causes airway hyperreactivity, eosinophilia, and goblet cell hyperplasia independent of the adaptive immune system. Separately, adaptive T cell-dependent type 2 airway inflammation was induced by intratracheal house dust mite (HDM) sensitization and challenge[210, 311]. Mice with or without pulmonary type 2 inflammation were then systemically infected with Staphylococcus aureus (S. aureus), which causes lethal sepsis and is commonly isolated from blood cultures of septic patients[209, 320]. Greater than 50% of control mice receiving intratracheal phosphate-buffered saline (PBS) prior to infection with S. aureus died by day 3, with 85% dead by day 5 (Figure 2.26). However, pre-treatment with IL-33 to induce innate type 2 inflammation significantly protected mice from mortality, with 50% of mice still alive by day 7. More dramatically, HDM sensitization and challenge to induce adaptive Th2 responses prior to infection robustly protected against S. aureus-induced mortality, with 75% of mice still alive by day 7. Thus, the presence of pre-existing type 2-biased immune response protected mice from becoming septic and dying, thereby confirming our hypothesis generated from the insurance claims analysis. Our "big data" analysis allowed us to infer a novel disease mechanism with significant implications for our understanding of sepsis pathophysiology. Our confirmation of this mechanism using two different mouse models establishes a new paradigm for translational sepsis research in which analysis of patterns in administrative data generates hypotheses to be tested using reductionist mouse approaches. Our findings demonstrate that an individual's immune status prior to infection has a significant impact on their risk for the development of sepsis and may also impact their disease course and outcomes. During sepsis, activation of type 1 or type 17 inflammatory responses leads to the robust production of pro-inflammatory cytokines like IL-6, TNF-$\alpha$, IFN-$\gamma$, and IL-17A2. The "cytokine storm" of this response mediates cardinal features of septic inflammation, including phagocyte recruitment to the site of infection, pathogen elimination, and inadvertent host tissue destruction. In addition,

these same type 1/type 17 pathways are aberrantly activated in the immune diseases that we found confer an increased risk of a diagnosis of sepsis, supporting the importance of pre-existing immune diseases on the risk for sepsis[305, 55]. The surprising finding that type 2 diseases were underrepresented in our septic patients suggests a protective effect on sepsis development and/or outcomes. Type 2 immune responses can antagonize pro-inflammatory type 1 and 17 responses and are often considered anti-inflammatory and tissue reparative[91]. For instance, in our mouse models, we induced innate type 2 responses via activation of pulmonary type 2 innate lymphoid cells (ILC2s), and induced adaptive type 2 responses via HDM sensitization/challenge to activate Th2 lymphocytes; both ILC2s and Th2s in the lung produce amphiregulin, an epithelial growth factor that restores tissue integrity following injury associated with inflammation[322]. Understanding the role of innate and adaptive type 2 responses prior to and during septic inflammation will inform the development of improved risk-prediction algorithms and reveal novel therapeutic targets. Indeed, therapeutic helminth infection in mice to activate type 2 responses may improve the outcomes of bacterial sepsis, and conversely, the explosion of anti-type 2 therapies may have the unintended consequence of potentiating sepsis morbidity and mortality in asthma patients[111, 132]. We did not adjust for medication use and disease control in our analysis, however, nearly all of the diseases listed in Table 1 are treated with local or systemic immunosuppression, thereby reducing the significance of this as a confounder. The pathophysiology of many of the listed immune-mediated diseases is not fully understood, and there are well-documented limitations in using administrative billing data for identification of patients with a specific disease[48]. Our mouse models were neither subjected to immunosuppressive or antimicrobial medications, nor do they have the genetic and comorbid illness heterogeneity observed in humans. Nevertheless, the clustering of only type 2 diseases as protective in the claims dataset and the robust protection noted in the mouse models indicates mechanistic specificity. In summary, we combined analysis of large-scale administrative data with analysis of mouse models to reveal a novel, unappreciated immunologic disease mechanism in sepsis.

Consideration of baseline immunologic bias as manifested by comorbid illnesses may predict hospital course and outcomes among acutely infected patients, and modulating type 2 responses could conceivably improve those outcomes, a possibility that remains to be tested.

**Table 2.8: Odds of immune-mediated diseases among septic and non-septic patients.**[a] diseases considered type 2-mediated. P value (mtc) refers to the multiple test comparison corrected p value, with significant at p¡0.05 and "ns" = not significant

| | **Septic** | **Not septic** | **Odds ratio (CI)** | **P value (mtc)** |
|---|---|---|---|---|
| Males, total n (%) | 36,594 (100) | 182,970 (100) | | |
| Sjogren's syndrome | 42 (0.11) | 460 (0.25) | 0.456 (0.410 – 0.557) | 1.70E-04 |
| [a]Food allergy | 187 (0.5) | 1,942 (1.1) | 0.479 (0.410 – 0.557) | 1.48E-22 |
| [a]Allergic rhinitis | 4,046 (11.1) | 37,244 (20.4) | 0.486 (0.470 – 0.528) | 5.18E-296 |
| Graves disease | 131 (0.4) | 976 (0.5) | 0.670 (0.553 – 0.804) | 0.016 |
| [a]Asthma | 4,027 (11.0) | 28,027 (15.3) | 0.684 (0.660 – 0.708) | 2.73E-103 |
| [a]Atopic and contact dermatitis | 5,316 (14.5) | 34,115 (18.7) | 0.742 (0.719 – 0.765) | 6.86E-78 |
| Celiac disease | 104 (0.3) | 694 (0.4) | 0.749 (0.603 – 0.921) | ns |
| Sarcoidosis | 253 (0.7) | 1,408 (0.8) | 0.898 (0.781 – 1.028) | ns |
| Dermatomyositis and polymyositis | 114 (0.3) | 585 (0.3) | 0.974 (0.790 – 1.193) | ns |
| Lupus erythematosus | 296 (0.8) | 1,457 (0.8) | 1.016 (0.893 – 1.152) | ns |
| Myasthenia gravis | 84 (0.2) | 410 (0.2) | 1.024 (0.800 – 1.300) | ns |

Continued on next page

**Table 2.8 – continued from previous page**

| | Septic | Not septic | Odds ratio (CI) | P value (mtc) |
|---|---|---|---|---|
| Type 1 diabetes mellitus | 5,604 (15.3) | 22,490 (12.3) | 1.290 (1.250 – 1.332) | 2.22E-50 |
| Vasculitis | 597 (1.6) | 2,222 (1.2) | 1.349 (1.230 – 1.478) | 8.51E-07 |
| Ulcerative colitis | 934 (2.6) | 3,317 (1.8) | 1.419 (1.317 – 1.527) | 4.63E-16 |
| Multiple sclerosis, other demyelinating diseases | 601 (1.6) | 1,808 (1.0) | 1.673 (1.522 – 1.837) | 1.12E-21 |
| Kawasaki disease | 30 (0.08) | 54 (0.03) | 2.780 (1.717 – 4.422) | 0.044 |
| | | | | |
| Females, total n (%) | 36,993 (100) | 184,965 (100) | | |
| [a]Allergic rhinitis | 5,698 (15.4) | 54,005 (29.2) | 0.442 (0.429 – 0.455) | 7.99E-269 |
| [a]Food allergy | 281 (0.8) | 2,741 (1.5) | 0.509 (0.448 – 0.576) | 3.01E-28 |
| Sjogren's syndrome | 363 (1.0) | 2,958 (1.6) | 0.610 (0.545 – 0.681) | 1.44E-17 |
| Graves disease | 357 (1.0) | 2723 (1.5) | 0.652 (0.582 – 0.729) | 4.70E-12 |
| Celiac disease | 186 (0.5) | 1,378 (0.8) | 0.673 (0.574 – 0.786) | 2.75E-04 |
| [a]Atopic and contact dermatitis | 6,729 (18.2) | 45,297 (24.5) | 0.686 (0.666 – 0.705) | 1.73E-153 |

Continued on next page

**Table 2.8 – continued from previous page**

| | Septic | Not septic | Odds ratio (CI) | P value (mtc) |
|---|---|---|---|---|
| [a]Asthma | 6,484 (17.5) | 42,386 (22.9) | 0.715 (0.694 – 0.736) | 1.29E-116 |
| Sarcoidosis | 429 (1.2) | 2,155 (1.2) | 0.995 (0.895 – 1.105) | ns |
| Myasthenia gravis | 116 (0.3) | 555 (0.3) | 1.045 (0.848 – 1.279) | ns |
| Systemic lupus erythematosus | 1,515 (4.1) | 6,286 (3.4) | 1.214 (1.146 – 1.286) | 1.47E-07 |
| Dermatomyositis and polymyositis | 214 (0.6) | 839 (0.5) | 1.277 (1.093 – 1.486) | ns |
| Vasculitis | 926 (2.5) | 3230 (1.7) | 1.445 (1.340 – 1.556) | 9.35E-18 |
| Multiple sclerosis,other demyelinating diseases | 1,163 (3.1) | 3,959 (2.1) | 1.484 (1.388 – 1.586) | 5.57E-26 |
| Ulcerative colitis | 1,140 (3.1) | 3,802 (2.1) | 1.515 (1.415 – 1.621) | 3.74E-28 |
| Type 1 diabetes mellitus | 5,184 (14.0) | 15,927 (8.6) | 1.730 (1.672 – 1.789) | 8.87E-205 |
| Kawasaki disease | 25 (0.07) | 31 (0.02) | 4.034 (2.283 – 7.061) | 1.61E-03 |

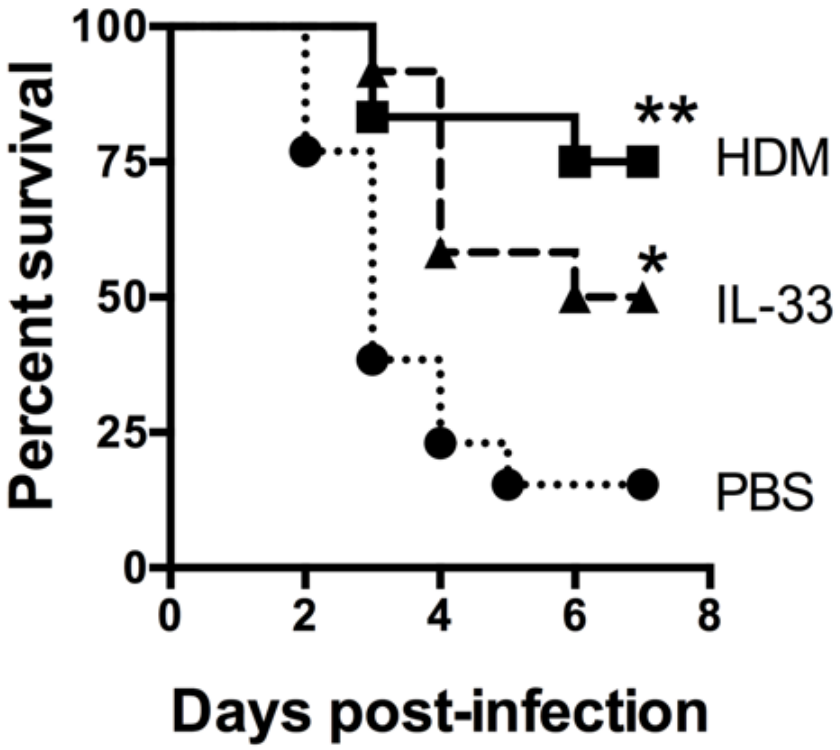


**Figure 2.26: Pre-existing type 2 immune responses protect against S. aureus mediated mortality in mice.** C57BL/6 mice were administered intratracheal IL-33 or HDM prior to being intravenously infected with a lethal dose of S. aureus USA 300. Controls received PBS. Shown are 2 pooled experiments for each group, with a total of 12-13 mice per group. PBS vs IL-33 *p=0.01; PBS vs HDM **p=0.001

### *2.5.2 Methods*

The Truven Marketscan Claims and Encounters database contains US insurance claims data, including International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes for inpatient and outpatient encounters, as well as other patient-level data including age, gender, and location, over the time period 2003-2013. We performed a case-control analysis of the Marketscan database as follows: all patients having either of 2 diagnostic ICD-9-CM codes for sepsis or severe sepsis (995.91 or 995.92, respectively) were identified as cases; controls were then matched to the cases in a ratio of 5:1 on age, geographic location, gender, and number of comorbid diagnoses (as a surrogate for patient complexity).

Both groups were analyzed for the presence of various comorbid immunologic diagnoses (as follows) occurring at any point during the time period covered in the database, in either inpatient or outpatient encounters:

Allergic rhinitis: 477, 477.0, 477.2, 477.8, 477.9

Asthma: 493, 493.0, 493.00, 493.01, 493.02, 493.1, 493.10, 493.11, 493.12, 493.2, 493.20, 493.21, 493.22, 493.8, 493.81, 493.82, 493.9, 493.90, 493.91, 493.92

Atopic and contact dermatitis: 373.32, 691, 691.8, 692, 692.0, 692.1, 692.2, 692.3, 692.4, 692.6, 692.8, 692.81, 692.83, 692.84, 692.89, 692.9, 693, 693.0, 693.8, 693.9, 708, 708.0, 708.1, 708.2, 708.3, 708.4, 708.5, 708.8, 708.9

Celiac disease: 579

Dermatomyositis and polymyositis: 710.3, 710.4

Food allergy: 477.1, 558.3, 692.5, 693.1

Graves disease: 242.0, 242.00, 242.01

Kawasaki disease: 446.1

Lupus erythematosus: 373.34, 695.4, 710.0

Multiple sclerosis, other demyelinating diseases: 340, 341.0, 341.1, 341.9

Myasthenia gravis: 358.0. 358.00, 358.01

Sarcoidosis: 135

Sjogren's syndrome: 710.2

Type 1 diabetes mellitus: 250.03, 250.01, 250.11, 250.13, 250.21, 250.23, 250.31, 250.33, 250.41,

250.43, 250.51, 250.53, 250.61, 250.63, 250.71, 250.73, 250.81, 250.83, 250.91, 250.93

Ulcerative colitis: 556, 556.0, 556.1, 556.2, 556.3, 556.4, 556.5, 556.6, 556.8, 556.9

Vasculitis: 446, 446.0, 446.2, 446.20, 446.21, 446.29, 446.3, 446.4, 446.5, 446.6, 446.7, 447.6

While the choice of immune-mediated diseases was arbitrary, a number of immunologic diseases were specifically excluded from the results because they were either congenic (and therefore present throughout the life of the patient) or have a poorly understood patho-

physiology. Documentation of a comorbid immune-mediated disease need not have occurred simultaneously with the diagnosis of sepsis; rather, patients were considered to have a comorbid immunologic disease if they had ever had such a diagnosis during the time period captured in the Marketscan database. Once the prevalence of these comorbid immune conditions was determined, the odds ratios were calculated between septic and non-septic groups and compared using Fisher's Exact Test. P-values for significant associations were adjusted for multiple test comparisons (mtc) using the Benjamini-Hochberg correction with a false discovery rate threshold of 0.1%, with the resultant P-value mtc ¡0.05 considered significant[21]. Data are expressed both as absolute numbers and as percentages of the total in parenthesis. Odds ratios are expressed with their confidence intervals, determined prior to multiple test correction. To conduct animal studies, C57BL/6 mice were purchased from Jackson Laboratories and maintained in house. Only female mice 6 to 7 weeks old weighing 14-17 g were used. All animal procedures and protocols were approved by the University of Chicago Animal Resources Center. The studies conformed to the principles set forth by the Animal Welfare Act and the National Institutes of Health guidelines for the care and use of animals in biomedical research. To induce an innate type 2 pulmonary response, 100 ng of recombinant murine IL-33 (eBioscience) or PBS control was intratracheally administered to anesthetized mice for 3 days. On day 4, mice were infected with 5x107 CFU of S. aureus USA300 LAC/100 $\mu$l via retro-orbital injection as previously described[229]. To induce an adaptive type 2 pulmonary response, mice were intratracheally sensitized with 100 $\mu$g of HDM (Greer Laboratories) or PBS control on day 0 and challenged with 25 $\mu$g HDM on days 7, 8, 9 and 107. Flow cytometry analysis of lungs from select mice confirmed the induction of type 2 pulmonary immune responses[299]. On day 14 mice were infected with 5x107 S. aureus as above. For both experiments, mice were euthanized when they reached 70% starting weight according to protocol or on day 7. Statistical analysis was performed using GraphPad Prism 6. Survival curves were analyzed using the Log-rank test. Significance was determined at $p<0.05$.

## 2.6 Acknowledgments

For the first section of this chapter, Andrey Rzhetsky and Robert Gibbons conceived and designed the experiments. Andrey and myself performed the experiments and analyzed the data. Andrey Rzhetsky, Christopher Lyttle, Steven Bagley and myself contributed analysis tools. Steven Bagley, Edwin Cook, Russ Altman, Robert Gibbons and Andrey Rzhetsky contributed and edited the text. The work described within this section was published [244]

For section 3.2, Joseph Lykins, Christopher Lyttle and myself performed the experiments, analyzed the data and wrote the text. Andrey Rzhetsky and Rima Mcleod conceived and supervised the project, helped design and implement the analyses, and wrote the text. Kelsey Wheeler, Fatima Clouser, Ashtyn Dixon, Kamal El Bissati, and Ying Zhou also contributed. The work described within this section was published [179]

For section 3.4, Paulette Krishack designed and implemented the mouse model experiments. I designed and implemented the computational experiments. Andrey Rzhetsky, Julian Solway, Anne Sperling, and Philip Verhoef conceived and supervised the project, helped design, implement the analyses, and edit the text. A revised version describing the work within this section was accepted for publication [157]

# CHAPTER 3

# QUANTIFYING GENETIC AND ENVIRONMENTAL CONTRIBUTIONS ACROSS MULTIPLE DISEASES

## 3.1 Introduction

Disease classifications (nosologies) are used ubiquitously in academic medicine, human genetics, the health industry, and economics. Much like any library's content catalogue, disease taxonomies strive to group together similar entities for ease of access and analysis. Initially, many of these groupings were largely arbitrary—often guided by topographical, anatomical, or even cultural similarities.[295, 148]

Historically, changes in these groupings have reflected a progression towards etiologic, common-cause disease classifications.

The evolution of nosologies has closely paralleled the evolution of methods designed for reconstruction of the universal tree of life. Approaches to species classifications were initially subjective, heuristic,[70, 57, 300, 283, 6] and made without any hint of the common-origin interpretation, utilizing only a small subset of all visible morphological features of any given organism. These early phylogenetic methods were followed by the use of maximum parsimony methods, explicitly minimizing the number of differences between proximal taxonomy leaves. Most recent arrivals to phylogenetics are statistical tree-making methods,[79] which infer taxonomies from very large datasets using explicit stochastic models of diverging organism traits during speciation.

In this study, we synthesized a synergy of the analytical methods developed for phylogenetic analysis with those established in dissecting the heritable and environmental components of human disease. The main premise of our analysis was that etiological disease taxonomy can and should be constructed using the explicit and objective genetic and environmental correlations between diseases.[278] Such a classification would maximize genetic and/or environmental disease similarities that have clustered together and would generate

the closest yet approximation to the common-cause nosology.

Our study used a dataset summarizing health information for more than one-third of the U.S. population, including more than 40 million families. The most informative subset of these, 481,657 unique individuals grouped into 128,989 families, was chosen for in-depth genetic analysis. In this study, we focused on estimating heritability, and environmental and genetic correlations between common diseases that were unambiguously encoded in the insurance claims. Doing so, we were unable to analyze quantitative traits, which are not represented in insurance claims.

A trait's narrow-sense heritability is defined as the ratio of its additive genetic variance[81, 317] to its total phenotypic variance see [181], p. 170. The environment-related counterpart to narrow-sense heritability is, consequently, the ratio of the environment-related variance (shared by siblings, parents, or the entire family) to the total disease-specific phenotypic variance. The environment-related variance portion of this ratio can be called preventability because it indicates the putative efficacy of interventions via changing environmental conditions.

## 3.2 Results

### *3.2.1 Data*

Our dataset was generated by subsampling from a very large collection of families represented in a compilation of insurance claims from Truven MarketScan. By definition, the dataset included only information about insured families, and therefore it is slightly biased towards more affluent urban populations, see Figure 3.1A. The largest families, as well as the majority of all families, were urban, see Figures 3.1A and B, with the overall urban population share slightly higher than 80.7% reported by US Census.[96] It is therefore unlikely that our heritability and genetic correlation estimates were affected by the sampling of families from rural areas, where average relatedness of individuals in the same county is potentially

higher than the country average.

The need to focus on a subset of families out of the total 40 million families was two-pronged. First, computational tractability demanded that we significantly restrict the sample: the bivariate analysis of common diseases can become impractical if the sample is too large. Second, in insurance claim, the data of parents and children are linked for a limited time, typically until children leave their parental insurance policy before age of 30, see Figure 3.1C. Therefore, we focused on a set of 128,989 families where both parents and children were "visible" for the longest time interval. No individual in the data was "visible" for more than 10-years.

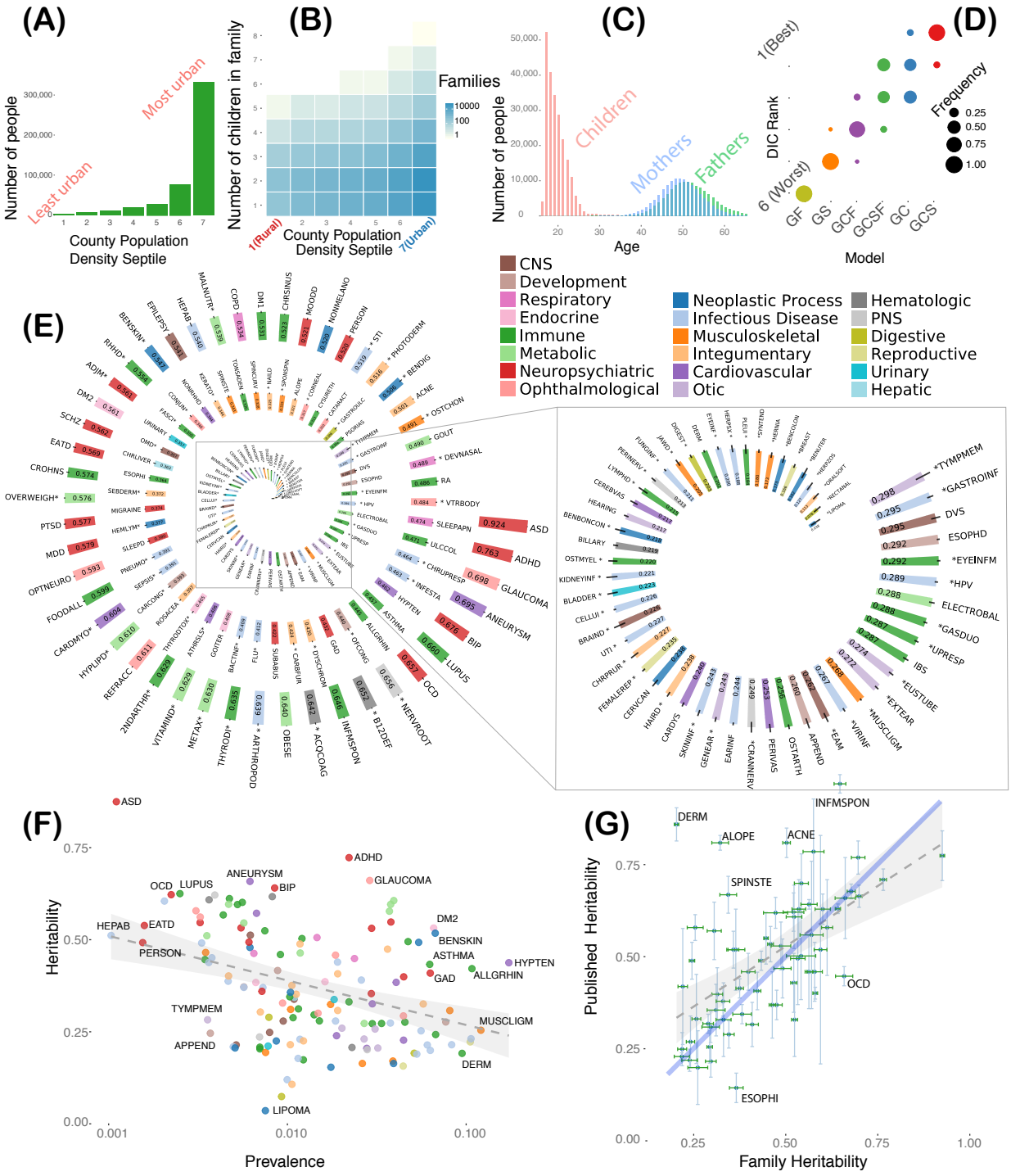


**Figure 3.1: Information on study population, results of model selection, and analysis of heritability of 149 diseases.**

**Figure 3.1 Continued** (A) Distribution of study population across population density septile; septile 1 corresponds to the most rural counties and septile 7 corresponds to the most urban counties. (B) Number of children in a family as a function of population density septile; septile notations are the same as in the (A). (C) Bar plot of parent/child age distribution in families described in our dataset. (D) Results of model selection, using univariate models GF, GS, GCF, GCSF, GC, and GCS, where G stands for additive genetics, F for common family environment, S for common sibling environment, and C for environment common for parental couple; plot shows frequency of corresponding model becoming the "best" (rank 1) as compared by DIC, second best (rank 2), and so on; clearly, the GCS model wins in the majority of cases. (E) Disease heritability estimates with one standard deviation; diseases, heritability for which appears to be measured for the first time are marked with asterisk; heritability values are sorted in decreasing order; color of the bar indicates biological system associated with the disease, see key in the upper right corner of the plate; keys to disease acronyms are given in the Supplement Table A.1 and A.1. (F) Estimates of disease heritability values against estimates of disease prevalence; the linear correlation is significantly negative, Pearson's r = –0.212 (95% CI [–0.36 –0.05]), and p = 0.00915. (G) Comparison of our estimates of heritability with the previously published estimates; see Supplement Table A.6 for detailed numbers; as expected, the estimates are highly correlated, but not identical, Pearson's r = 0.571, CI (0.380, 0.715), p = 6.90 10-7. Some of the differences are due to variation in the breadth of definition of a disease, for example, our definition of atopic contact dermatitis (DERM) is much broader than the disease definition in the comparison study.

### *3.2.2 Model selection*

We started our analysis with a systematic comparison of those mathematical models most likely to describe the structure of phenotypic variance of the families in our dataset (see Figure 3.1D; DIC stands for Deviance Information Criterion commonly used in Bayesian

model selection [95]). The best model included shared couple (parents) environment, shared sibling environment, and additive genetics (GCS model in the Figure 3.1D). The second-best model dropped the shared-sibling environment component, S, see Figure 3.1D. We then used the GCS and GC models (whichever fit data best) to estimate heritability and preventability for 149 common diseases, see Figure 3.1E and Supplementary Figure A.1. (Disease abbreviations are spelled out in Supplementary Table A.1.)

### *3.2.3 Estimates of heritability*

We estimated the narrow-sense heritability for 149 of the most common diseases present in the insurance claim dataset, Figure 3.1E. To the best of our knowledge, these estimates were obtained for the majority of diseases for the first time in our study: 84 out of 149 estimates (56 percent) are new. These putative first-time estimates are marked with asterisks in Figure 3.1E (see details in the Supplementary Table A.1).

Our heritability estimates spanned a wide range of values, from 0.924 (autism) to 0.038 (lipoma). Reasoning from a theoretical equation that links disease prevalence to its relative risk in siblings and its heritability,15 there would be an expectation that, on average, diseases with higher heritability would be more prevalent than those with lower heritability. On the other hand, rarer diseases could be, on average, more homogeneous with respect to both genetic and non-genetic etiologies than more common diseases, which might preclude seeing a clear relationship. In reality, the apparent correlation between our estimates for heritability and disease prevalence turned out to be significantly negative (Figure 3.1F): The estimated linear regression slope was – 1.20 (se = 0.455, p = 0.00915), with Pearson's r = –0.212 (95% CI [–0.36 –0.05]), and p = 0.00915.

Out of the 65 diseases with previously published heritability estimates in our disease set, 52 of our estimates agreed with the published estimates within 95 percent CI (see Figure 3.1G). The published and new estimates were highly correlated (r = 0.571, CI (0.379, 0.715), p = 6.902 10-7, linear slope 0.4975, se = 0.0902, p = 6.90 10-7). Furthermore, the error bars

for the new heritability estimates (Figure 3.1G) are predominantly much narrower than those published. The mean values of our heritability estimates are, on average, slightly lower than previously published values, as can be seen by comparing the dotted regression line (slope = 0.5) with the blue line (slope = 1) in Figure 3.1G. Various possible sources of this trend have been enumerated in the Discussion.

According to common genetics wisdom, diseases with early onset tend to have higher heritability. This assumption was tested by using heritability and onset estimates of our 149 chosen diseases (see Supplemental Figures A.2 A and B). The correlation between the age of onset and disease heritability appears negative for a subset of diseases, including those currently categorized as neuropsychiatric, neoplastic, metabolic, ophthalmologic, and central nervous system diseases. For diseases with strong immune system component, such as autoimmune and infectious diseases, the estimated correlation between heritability and disease onset was positive (see Supplementary Figure A.2B). When combined, the heritability estimates for all diseases, contrary to the common wisdom, showed no linear correlation with age of disease onset.

Our analysis also provided estimates of the environmental counterparts of heritability: unique-environment, common-couple, and common-sibling preventability (see Supplemental Figure A.1A, B, and C). The common-couple preventability estimates range from 0 (autism) to 0.46 (photo dermatitis); the corresponding common-sibling estimates tend to be smaller, but can be as large as 0.29 (sepsis). The estimates for unique-environment preventability tended to be the largest: in our dataset, estimates ranged from 0.03 (eye infection) to 0.842 (diseases associated with damages to rectum and anus). For example, the largest preventability estimate for migraine is for unique environment (0.534), followed by common-couple (0.11), and negligible common-sibling preventability. Similarly, for sleep disorders, preventability estimates were 0.269, 0.22, and 0.15 for unique, couple and sibling preventability, respectively.

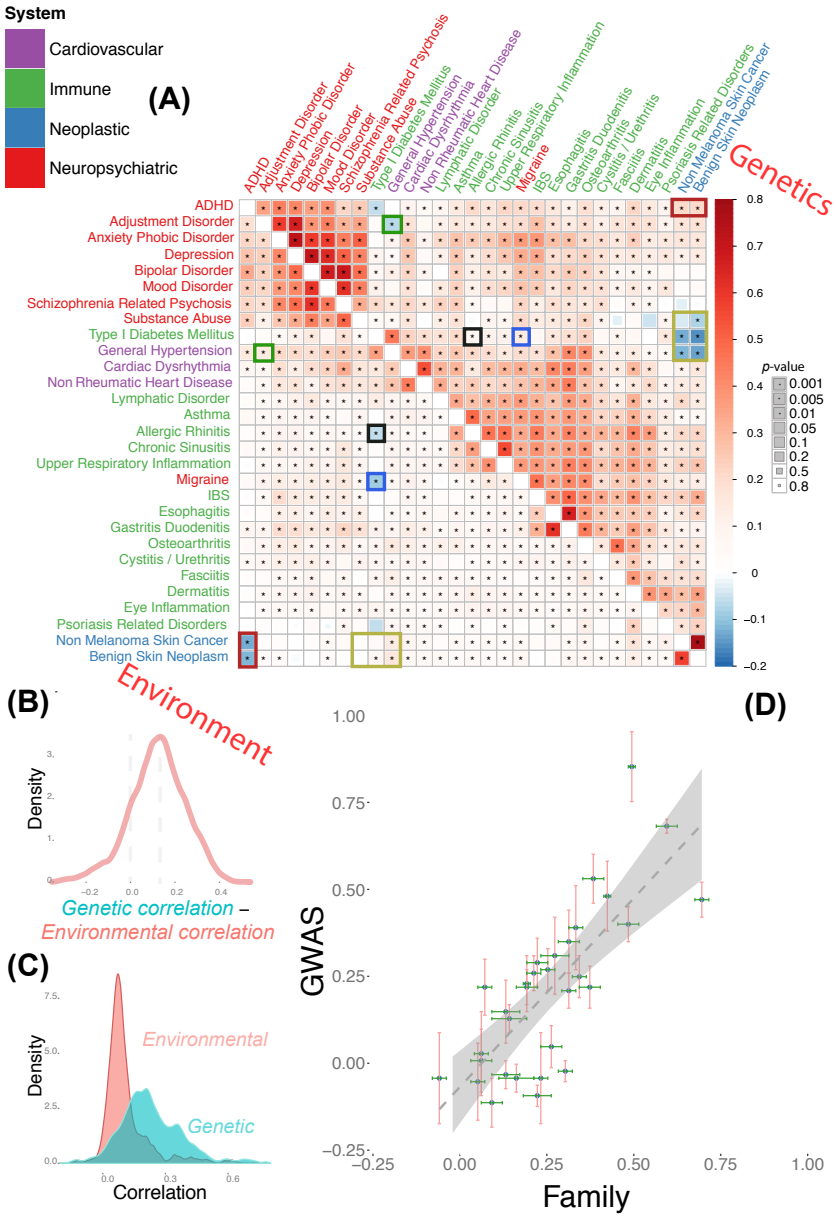


**Figure 3.2: Genetic and environmental correlations between diseases.**

**Figure 3.2 Continued** (A) Matrix of pairwise genetic correlations (upper half) and corresponding environmental interactions (lower half of the matrix) colored by sign and magnitude (see legend) The disease color labels indicate biological systems associated with a particular disease; the size of the squares indicates statistical significance of the correlation, see key on the right. Cells with asterisks indicate pairwise interactions that remained significant at a false discovery rate of 1 percent.[] The color boxes within the matrix indicate opposite-sign correlation values for the same pair of diseases. Posterior probabilities of two correlation values (genetic and environmental) for the same pair of diseases having the same sign were 1.869 $\times 10^{-14}$ (ADHD and benign skin neoplasm), 3.376 $\times 10^{-14}$ (ADHD and non-melanoma skin cancer), 4.523 $\times 10^{-9}$ (adjustment disorder and general hypertension), 8.715 $\times 10^{-4}$ (migraine and type 1 diabetes mellitus), 9.251 $\times 10^{-5}$ (benign skin neoplasm and type 1 diabetes mellitus), 6.401 $\times 10^{-33}$ (benign skin neoplasm and general hypertension), 3.712 $\times 10^{-17}$ (non-melanoma skin cancer and general hypertension), 3.933 $\times 10^{-4}$ (allergic rhinitis and type 1 diabetes mellitus). (B) Distribution of (Genetic correlation - Environmental correlation) values for the same pair of diseases. (C) Individual distributions of genetic and environmental correlations superimposed on the same plot. (D) Comparison of our family-based estimates of genetic correlations between diseases compared to previously published GWAS-based values, the complete data on values and references is provided in the Supplement Table A.5. Linear fit with a slope of 1.08 (SE=0.167) is indicated by the dotted line.

### *3.2.4 Genetic and environmental correlations*

Our analysis of pairwise disease correlations focused on 29 diseases, of which all pairs were well-represented in both the children and parents of our dataset (see Supplemental Table A.1). We estimated genetic and environmental correlations across all pairs of these 29 diseases (Figure 3.2A-D, Supplemental Table A.4).[107] The majority of correlation values in our analysis differed significantly from zero (the null hypothesis r = 0) at a 1 percent false

discovery rate[21]. On average, genetic correlations between diseases tended to be stronger than their corresponding environmental correlations (see Figures 3.2B and C). However, for the majority of neuropsychiatric disease pairs, the environmental correlations are nearly as strong as the genetic correlations. In some cases, such as for the substance abuse and schizophrenia disease pair, the environmental correlation is stronger than the genetic correlation, and nearly equal for other disease pairs, such as schizophrenia and bipolar disorder. This observation is consistent with an earlier finding of nearly equal amounts of shared genetic and environmental effects between schizophrenia and bipolar disorder.[169]

Figure 3.2C indicates that the environmental correlation distribution has a longer positive tail than the more symmetric genetic correlation distribution. Genetic and environmental correlations for the same disease pair were themselves positively correlated, and genetic correlations were also positively correlated with phenotypic correlations (see Supplemental Figures A.3A and B).

In a few cases, direction of correlation was reversed between genetic and environmental components, indicated with color rectangles in Figure 3.2A; the corresponding Bayesian posterior probabilities for significance of sign difference are shown in the figure legend. These cases were particularly unexpected, as they indicate hypothetical scenarios where genetic and environmental factors act antagonistically in determining a phenotypic path bifurcated between two apparently unrelated diseases.

On average, family-based estimates of genetic correlations obtained in our study have much narrower error bars (with a few exceptions) than earlier genome-wide association study (GWAS) estimates, mostly due to the very large sample size of our dataset. It is quite remarkable that genetic correlations obtained by two different methods agree so well. GWAS genetic correlations and family study genetic correlations estimate different quantities: family studies estimate the correlation of the total genetic variation (both rare and common), while genetic correlations, estimated using single-nucleotide variants (SNPs), are based on genotyped and imputed common SNPs, which are only a subset of the total genetic

variation. Essentially, our data suggest that family-based estimates of genetic correlations reflect predominantly common variants.

The absolute values of genetic correlations are high for several common conditions across all the diseases that we analyzed (for example, asthma, allergic rhinitis, osteoarthritis, and dermatitis). This result is surprising as it suggests that the most prevalent complex diseases share a considerable amount of predisposing variation, even across apparently dissimilar diseases. Human genetic variation associated with common diseases appears highly pleotropic.

In order to get a baseline of the expectedness (or unexpectedness) of observed patterns in genetic and environmental correlations, we used the International Classification of Diseases version 9 (ICD-9) (see Supplemental Figure A.4, left). Based on the ICD9 taxonomy, genetic and environmental correlations for migraine are surprising. As migraine is clearly associated with the central nervous system, one would expect that its etiology is most similar to those of other neuropsychiatric conditions. For example, "Mental disorders," codes 290-319 in ICD9 taxonomy, have a sister group of "Diseases of central nervous system and sensory organs," codes 320-389, containing both migraine and eye inflammation. However, in our analysis of its genetic and environmental correlations, migraine is not similar to other nervous system diseases. Rather, it is much closer to immune system diseases, such as irritable bowel syndrome (IBS) in the genetic correlation space, and to cystitis/urethritis in the environmental correlation space. These findings suggest that migraine is associated with general, not nervous-system specific, inflammatory processes and can possibly be mitigated with some of the treatments that have been developed for inflammatory diseases.

Inferring nosologies from correlations: A logically consistent way to examine all similarities between disease pairs simultaneously is to transform the genetic and environmental correlations shown in Figure 3.2A into distances, and then infer objective genetic and environmental disease classifications from those distances. We chose to use the simplest (1 – correlation) distance transformation. The distance-matrix method that we used[247] for this purpose is designed to identify the classification topology that approximates the distance-

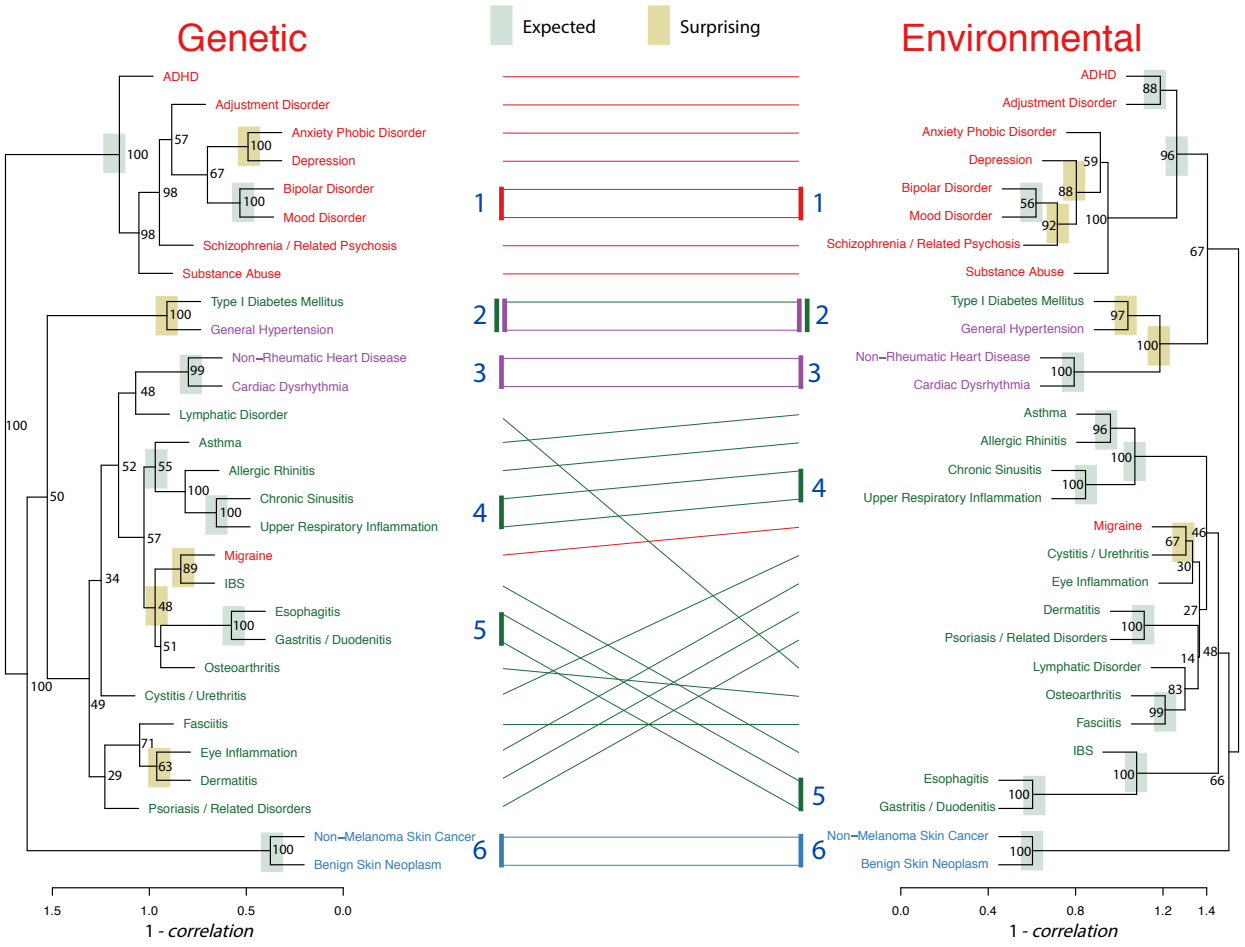


**Figure 3.3: Neighbor-joining classifications showing the 29 conditions' nosologies inferred from genetic and environmental correlations presented on the left and the right tree, respectively.** For both classifications, we defined the distance between diseases as 1—*correlation*. Because we estimated a posterior distribution for each correlation estimate, we were able to sample 10,000 distance sets using posterior distributions for pairwise correlations. For each of these samples, we estimated a classification and computed reliability measures for individual classification topology partitions (each integer number on the tree indicates the percent of trees out of 10,000 in which this particular partition was present). The disease labels are colored according to associated biological systems, consistent with other figures. Note that, while the genetic and environmental trees are significantly different, both are stable, as the bootstrap-like numbers indicate.

matrix the closest, so that the length of the shortest path connecting two classification leaves closely approximates distance in the input matrix. This method is not dependent on the assumption that all expected distances between leaves and root are equal. By repeatedly sampling distances from their posterior distribution, one can each time compute a tree from the resampled distances, counting the percentage of times each disease grouping occurs in the resampled trees 20-22 (see Figure 3.3A and B). The distances between diseases in a classification is meaningful. For example, two diseases connected with shorter branches are more tightly correlated and more similar genetically or/and environmentally than two diseases connected with very long branches. When a disease group is associ,ated with a reliability number of 100, it means that this particular disease partition was replicated in all trees. In other words, the bootstrap-like numbers [66, 79, 67] indicate the statistical reliability of the classification.

In this analysis, the bootstrap-like measures identified a number of remarkably stable disease clusters present in both genetic and environmental trees (Figure 3.3, clusters 1-6). We used the ICD9 disease taxonomy (Supplemental Figure A.4, left) as a baseline to identify disease groupings that are expected (based on ICD9 classification), and those that are unexpected (defiant of ICD9 classification, but statistically significant), see Figure 3.3A and B, green and yellow highlights, respectively. Many of the stable disease groups (with high bootstrap-like numbers) lie within the traditional view of disease similarity. However, many stable clusters defy the currently established nosology. For example, type 1 diabetes groups with general hypertension (support 96 and 100 in the environmental and genetic classifications, respectively)—these two diseases are not typically thought to be closely related (see Supplemental Figure A.4). Migraine in both inferred environmental and genetic classifications appears to be genetically similar to inflammatory diseases, such as irritable bowel syndrome (IBS).[101] In the ICD9 taxonomy, however, migraine is placed together with eye inflammation, in the cluster of diseases of the central nervous system and sensory organs. In our study's genetic tree, eye inflammation is far away from migraine, but is grouped with

dermatitis. In the environmental classification, migraine is the closest to inflammations associated with the infections cystitis and urethritis; eye inflammation is weakly grouped with the cluster of migraine-cystitis/urethritis. This suggests that migraine etiology is closely associated with immune system function and the established disease taxonomy needs revision.

Neuropsychiatric diseases stayed in the same stable cluster in both taxonomies in Figure3.3.[36] However, within the cluster, disease groupings varied considerably. In our genetic classification, depression was significantly grouped with anxiety. This is in contrast to ICD9 taxonomy, which places depression together with mood and bipolar disorders. In our environmental classification, schizophrenia is significantly closer to bipolar and mood disorders than to depression, again contrary to ICD9.

As expected, a classification computed from complete phenotypic correlations represents a compromise between genetic-only and environmental-only classifications (see Supplemental Figure A.4, right).

## 3.3 Discussion

We conducted a very large-scale, family-based, phenotypic-variance analysis of numerous complex diseases. Methodologically, our work is indebted to work of Lichtenstein et al. [169] and Xia et al. [318] in considering genetic data for nosology inference[169] and careful model selection.[318] It has been long suspected that complex diseases have numerous predisposing factors, both in the genetic and environmental realms. For the first time, we were able to compare the contribution of both environmental and genetic determinants to the phenotypic variances and covariances of a broad range of diseases, and transform these covariances into estimates of disease classifications.

Our study contributes to a series of influential, interlinked probes into complex disease heritability and cross-disease genetic correlations.[254, 35, 56, 52, 175] For example, Munoz et al. 2016 [198] studied 12 complex human diseases using 502,682 participants and the family histories of disease in 1.5 million individuals. As our dataset provides rich phenotypic

information on a very large population we were able to analyze heritability and preventability of a collection of diseases (149) an order of magnitude larger than what was previously done, using data for comparable number of individuals. Furthermore, the statistical power associated with this broad and complex sample provided a new opportunity to contrast genetic correlation estimates from family data with estimates that have been made using DNA variants. Confirming previous findings,[175, 298] we observed a near-linear relationship between common SNP- and family-based genetic correlations with a proportionality constant of 1.150 (se = 0.035, p ¡ 2 $\times 10^{-16}$) between total liability and genetic correlations. With an over one-fold increase in pairwise estimates and a much smaller standard error, our ratio estimate is within the confidence intervals of published results.[175, 298] These results suggest that the largest part of genetic correlation between complex diseases is associated with common variants captured by SNP genotyping.

As is true for most observational studies, there are several possible sources of biases. Family studies based on closely-related individuals may inflate narrow-sense heritability estimates due to unaccounted for effects of shared environment, maternal influences, or epistatic interactions of genetic variants. [171, 328] In agreement with previous findings regarding the significance of shared environmental effects, [318, 198, 323] our study provided first-time or updated heritability estimates for 149 diseases. On average, our heritability estimates were lower than those reported by twin/family studies by a factor of 0.90. Thus, we conclude that SNP-based heritability estimates explain, on average, 49 percent of our family-based heritability estimates, a 13 percent increase from previous estimates. As articulated by Zuk et al., [328] one of the major sources of bias in estimates of heritability is associated with the choice of mathematical model, as the narrow-sense heritability, by definition, does not account for potential deviations from genetic additive model. The insurance data describes, at best, 54.7 to 69.7 percent of the U.S. population, depending on age group,[38] so a considerable lower-income stratum of U.S. society is not represented in this dataset. Data from insurance claims does not include ethnicity and race; therefore, we were unable to explicitly

adjust for these confounders. The other contribution to the estimate bias can be attributed to assortative mating, as the US population is stratified by ethnicity, income, and geography, with all of these factors contributing to assortative marriages.

While ethnicity is not recorded in the US insurance claims, it can be imputed. For example, according to the US Centers for Disease Control and Prevention,[212] sickle cell disease (SCD) affects, on average, one out of every 365 African Americans; the incidence rate of SCD is about 88 times lower in the rest of the Americans. The incident rate for SCD in our database is 2.85523 10-4 vs 3.23891 10-4 in the nation on average. Given that the US African American population is 12.2 percent of the total [96, 293], African American patients appear to be represented in our data at 10.6 percent (about 13 percent lower than the average over the US). Given the very large sample size, the ethnic diversity of our dataset should be a reasonable representation of the multiethnic composition of the US insured population.

When computed solely from genetic information, "genetic correlation is immune to environmental confounding but is subject to genetic confounding." [35] In the case of family-based analyses, environmental confounding is an issue researchers might address with an appropriate mathematical model of genetic and environmental factors working in consort. Unfortunately, the appropriate model is unknown and, therefore, interpretation of results is conditional on the assumptions of a rather simple, additive-genetic and additive-environment model—a model that is used, in most studies, for lack of a better (experimentally-grounded) alternative. Another conceivable caveat is related to possible biases in the sampling of affected individuals. [35] Finally, our results reflect the medical coding of disease in the healthcare system rather than research-quality disease diagnoses. Extensive study of the correspondence in results of genetic association studies conducted with research diagnoses and those conducted using diagnoses from electronic health records have demonstrated good concordance for large association studies. [63]

Our results indirectly indicate that environmental patterns cooperate with genetic factors in triggering complex diseases in non-transparent combinatorial ways: environmental triggers

for genetically similar complex disorders are not necessarily similar, and vice versa. This observation raises very interesting and medically relevant questions such as "What is the correct nosology?" and "Should we classify conditions by the cause of the disease?" It is quite likely that, for complex diseases, the actual disease-predisposing factors are numerous, heterogeneous, and, possibly, combinatorial. Should we define myriad different diseases for every possible permutation of genetic and environmental determinants? The current practice of classifying diseases by the superficial similarity of their clinical signs is clearly suboptimal.

Lichtenstein et al.'s[169] study discussed the difficulties and ambiguities associated with changing uncertain diagnoses ("patients with one diagnosis sometimes evolve into the other diagnosis"), and, to a large extent, their discussion and hedging apply here. Type 1 and type 2 diabetes are excellent examples of this challenge. While type 1 diabetes leads to high blood glucose levels because of autoimmune destruction of the insulin-secreting beta-cells in the pancreas, and type 2 diabetes arises from complex metabolic dysregulation of insulin secretion, and the insensitivity of peripheral tissues to the action of insulin, the two are commonly conflated in electronic health records, even for the same patient. This is in part because their corresponding ICD9 codes are very close to one other, (250.01 and 250.02 for type 1 and 2, respectively), and in part because physicians often lack the data used in research studies to aid in making this distinction: some type 2 diabetic patients are inevitably classified as type 1 diabetics and, possibly, vice versa. This disclaimer regarding diagnostic uncertainty also applies to phenotypes with overlapping clinical signs, such as schizophrenia, schizoaffective and bipolar disorders, and depression, or benign and malignant skin cancers, but does not apply to diseases from radically different biological systems, such as schizophrenia and skin cancer.

## 3.4 Methods

### *3.4.1 Data*

Our study used data from the 2003-2011 Truven Health MarketScan Commercial Claims and Encounters Database, which comprised 115,805,687 individuals and 56,003,690 policies. We defined a family as a group of individuals on a single insurance policy. In each family, we assumed primary and secondary beneficiaries were parents and other dependents were children. We grouped ICD9 diagnostic codes into 568 categories based on their clinical manifestations. We then selected 29 complex diseases of four biological systems (neuropsychiatric, immune, oncological, and cardiovascular) for bivariate analysis and 149 diseases of 20 biological systems for univariate analysis based on disease prevalence (Supplementary Table A.1) with standard error calculated as:

$$\mathrm{SE} = \sqrt{\frac{p(1-p)(1-f)}{n}}$$

where $n$ is the total individuals and $p$ is the prevalence and $f$ is the fraction of the total US population sampled [293]. We calculated the phenotypes for each relational pair (parent-offspring, siblings, couple) and selected diseases with at least 30 data points per disease state and relational category. We calculated the age of onset for each disease as the five percent age percentile of all patients with a given disease in the database. To maximize the probability of correct genetic relatedness, we selected families with parents and dependents having at least 15 years' age difference. In addition, we set the minimum enrollment time to six years, ran our analysis using 128,989 nuclear families with the fullest medical history in our database, and with children aged 16 and above. The resulting 481,657 individuals had been enrolled in the database for an average of 6.5 years.

### *3.4.2 Statistical analysis*

We used a multivariate, generalized, linear mixed model with a probit link,[154] i.e., the probability for an individual to have a disease is measured by an underlying Gaussian latent variable (liability ) [73, 74]. This model [265] allows for inference of four kinds of factors influencing disease liability variation: genetic effects associated with pedigree, environmental effects shared by couples, environmental effects shared by siblings, environmental effects shared within families, and unique environmental effects.

The outcome vector $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2\}$ is a vector of case($\mathbf{y} = 1$) or control($\mathbf{y} = 0$) status for each disease on the observed scale. Let liability $\boldsymbol{\ell} = \Phi(\mathbf{y})^{-1}$, where $\Phi$ the Cumulative Distribution Function (CDF) of the standard normal distribution, and

$$\boldsymbol{\ell} = \mathbf{X}\beta + \mathbf{a} + \mathbf{c} + \mathbf{s} + \mathbf{f} + \mathbf{e}$$

where $\beta$ is the vector for fixed effects of age and gender, $\mathbf{a}$ is the vector of random additive genetic effects based on pedigree, $\mathbf{c}$ is the vector of environment effects shared by a couple, $\mathbf{s}$ is the vector of environment effects shared by siblings, $\mathbf{f}$ is the vector of environment effects shared by a family, and $\mathbf{e}$ is the vector of unshared unique effects. [265] We followed the naming convention used in [318].

The underlying binary traits' liability (co)variance structure [265] is: The (co)variance structure of the liability underlying binary traits is

$$\text{var}(\boldsymbol{\ell}) = \mathbf{G} \otimes \mathbf{A} + \mathbf{R_c} \otimes \mathbf{I_c} + \mathbf{R}_s \otimes \mathbf{I}_s + \mathbf{R_f} \otimes \mathbf{I_f} + \mathbf{R}_e \otimes \mathbf{I}_n$$

and $\mathbf{A}$ is the additive genetic relationship matrix, $\mathbf{I_c}$ is the couple environment matrix, $\mathbf{I_s}$ is the sibling environment matrix,$\mathbf{I_f}$ is the family environment matrix, $\mathbf{I_f}$ is an identity matrix, $\mathbf{G}$ is the additive genetic (co)variance, $\mathbf{R_c}$ is the couple environment (co)variance, $\mathbf{R_s}$ is the sibling environment (co)variance, $\mathbf{R_f}$ is the family environment (co)variance, and $\mathbf{R_e}$

is the unique environment (co)variance. Furthermore, individual covariance matrices are parametrized as

$$\mathbf{G} = \begin{bmatrix} \sigma^2_{a,1} & \sigma_{a,12} \\ \sigma_{a,12} & \sigma^2_{a,2} \end{bmatrix}, \mathbf{R_c} = \begin{bmatrix} \sigma^2_{c,1} & \sigma_{c,12} \\ \sigma_{c,12} & \sigma^2_{c,2} \end{bmatrix},$$

$$\mathbf{R_s} = \begin{bmatrix} \sigma^2_{s,1} & \sigma_{s,12} \\ \sigma_{s,12} & \sigma^2_{s,2} \end{bmatrix}, \mathbf{R_f} = \begin{bmatrix} \sigma^2_{f,1} & \sigma_{f,12} \\ \sigma_{f,12} & \sigma^2_{f,2} \end{bmatrix}, \mathbf{R_e} = \begin{bmatrix} \sigma^2_{e,1} & \sigma_{e,12} \\ \sigma_{e,12} & \sigma^2_{e,2} \end{bmatrix}$$

The narrow-sense heritability, couple environmental effects, sibling environmental effects, family environmental effects, and unique environmental effects for disease $j$ are defined on the liability scale in the following way;

$$h^2_j = \frac{V_{a,j}}{V_{P,j}},\ e^2_{c,j} = \frac{V_{c,j}}{V_{P,j}},\ e^2_{s,j} = \frac{V_{s,j}}{V_{P,j}},\ e^2_{f,j} = \frac{V_{f,j}}{V_{P,j}},\ e^2_{u,j} = \frac{V_{u,j}}{V_{P,j}},$$

The genetic correlation coefficient ($r_g$) and environmental correlation coefficient ($r_e$) were calculated as

$$r_g = \frac{\sigma_{a,ij}}{\sigma_{a,i}\sigma_{a,j}}, r_g = \frac{\sigma_{c,ij} + \sigma_{s,ij} + \sigma_{f,ij} + \sigma_{e,ij}}{\sqrt{\sigma^2_{c,ij} + \sigma^2_{s,ij} + \sigma^2_{f,ij} + \sigma^2_{e,ij}}\sqrt{\sigma^2_{c,ij} + \sigma^2_{s,ij} + \sigma^2_{f,ij} + \sigma^2_{e,ij}}}.$$

Due to the binary nature of our phenotypic data [97], we estimated variance components using Bayesian methods with the MCMCglmm package.[107] We used a chi-squared prior with one degree of freedom for the univariate analysis[59] and Half-Cauchy prior for the bivariate analyses.[94] For the univariate analyses, we ran a burn-in period of 150,000 to 330,000 iterations depending on convergence, and sampled 600,000 iterations with 500 thinning intervals. For bivariate analyses, we ran a burn-in period of 30,000 to 44,000 iterations, and sampled 120,000 iterations. We checked model convergence using both standard MCMC diagnostic

tests[93, 116, 225] and visual comparison after the burn-in period. We reported parameter estimations with posterior means, posterior standard deviations, and 95 percent confidence intervals (CI). The posterior distributions represent the distributions of true parameters, given the data and the priors. Posterior probabilities for sign differences between same disease genetic and environmental correlations were calculated assuming a bivariate normal posterior distribution. We corrected for multiple testing using the Benjamini-Hochberg method [21] and deemed a correlation significant if it passed the false discovery rate of one percent. We also constructed neighbor-joining trees based on a distance definition of 1-correlation for the correlation matrices[247]. We performed 10,000 simulations for each tree by sampling from the correlation posterior distributions. We calculated a bootstrap-like measure indicating the percentage of simulations that replicated the disease partition.

### *3.4.3 Model selection*

We conducted two rounds of model selection to find the most appropriate genetic and environmental models for both univariate and bivariate analysis using deviance information criteria (DIC). [267] The full model 'GCSF', as well as five simpler models, were selected based on 29 diseases involved in both univariate and bivariate analyses. We then conducted a second run of model selection between the top two models on all 149 diseases. Due to the high computational cost of bivariate analysis, we based our bivariate model on univariate model selections and chose 'GCS' model for the bivariate analysis.

### *3.4.4 Pedigree error*

Quantitative genetic estimations, such as those for heritability and genetic correlations, rely on the accuracy of the pedigree information. Intuitively, we expect a downward bias in both heritability and genetic covariances due to pedigree errors. Indeed, simulation and population studies have shown that heritability estimates were underestimated, albeit slightly; pedigrees with 20 percent errors led to five percent underestimation of heritability estimates.

[24, 41] Genetic correlation estimates were influenced even less by mis-assigned relations: Both [196] and [24] found no biases caused by pedigree errors in genetic correlation estimations using both simulated and real data. Stepchildren and adopted children We collected US Census data on children by household types.[156, 291] The 2010 US Census surveyed a large population and reported data for children of differing age groups, shown in Supplementary Table A.7. Supplementary Table A.8 is based on US Current Population Survey data from 2007 to 20011 for children under age 18. This data showed that the percentages of children living with both biological parents were consistent with percentages from US Census data. Pedigree simulation Following the simulation model from [41], we performed 100 simulations on 5000 nuclear families with 2.4 percent adoptive children and 6.2 percent stepchildren[156] and estimated parameters with the true pedigrees versus mis-assigned pedigrees. We used a stochastic simulation model to generate pedigrees of two generations, with varying heritability estimates (0.03-0.97) and genetic correlations (0.13-0.85). The parents are assumed to be unrelated and unselected individuals. We simulated two binary traits, following the model $y = I(l > 0), = \mu + Za + e$ where $y$ is the matrix containing individual phenotypes at both traits, is the underlying liability, $l$ contains the population means for liability, a is a matrix of additive genetic effects, and e is a matrix of residual errors. Z represents an incidence matrix of the individual effects a has upon liabilities in . All models were solved using the MCMCglmm. Indeed, we also found a mean underestimation of 5.6 (SE=0.56) percent for heritability and no evidence of biases for estimations of either genetic (t-test p=0.8784) or environmental correlations (t-test p=0.9948). We then calculated and reported heritability estimates adjusted for the underestimation.

### *3.4.5 Heritability comparison*

We compared our heritability estimates with results from other independent studies. We collected reference family heritability estimates from 65 out of the 149 traits we studied. We also collected 31 GWAS heritability estimates from literature. Most of the data is collected

through other sources; the comparisons are listed in Supplementary Table A.3 and A.6.

### *3.4.6 GWAS and family-based genetic correlations*

We compared our genetic correlation estimates with estimates using GWAS data on common pairs of traits. First, we collected genetic correlations from literature. [7, 35, 52, 175] Next, we compared those genetic correlation estimates we found in common. To maximize this comparison, we broadened the collection of traits to include non-rheumatic heart disease as a proxy for cardiovascular diseases [175] , and type I diabetes as a proxy for fasting glucose, see [224] for justification of this choice. The resulting 30 genetic correlation pairs (Supplementary Table A.5) showed a correlation of 0.769, 95 percent CI (0.571-0.883) between our estimates and GWAS results, along with a linear fit with a proportionality constant of 0.108 (SE=0.167), indicating consistency between the two methods.

### *3.4.7 Heritability, disease prevalence, and sibling relative risk*

To approximate heritability estimates from disease prevalence and sibling relative risks, we used the following equations adopted from [316]:

$$h_L^2 = \frac{2[T - T_1\sqrt{1-(T^2-T_1^2)(1-T/i)}]}{i + T_1^1(i-T)}$$

where $K$ is disease prevalence, , $T = \Phi^{-1}(1-K)$ , $T_1 = \Phi^{-1}(1-\lambda_s K)$ and $\Phi$ is the standard normal CDF, is the sibling relative risk, $i = z/K$, and $z$ is the height of the standard normal curve. This model assumes only additive genetics and unique environmental effects and represents the heritability on the liability scale.

## 3.5 Acknowledgments

Andrey Rzhetsky designed experiments, analyzed data, and wrote the text with myself. Hallie Gaitsch and Hoifung Poon helped perform computational experiments. Hallie Gaitsch, Nancy Cox, Hoifung Poon contributed to iterative improvement of the text. A revised version describing the work within this Chapter was published [303].

# CHAPTER 4

# ANNOTATING TEXTUAL CORPUS FOR DIGITAL PHENOTYPING

## 4.1 Introduction

**Background** Natural language processing and text mining are essential for extracting information from biomedical literature, scientific research articles, and clinical text. The development and evaluation of such tools are limited by the availability of carefully annotated biomedical corpora. Particularly, information extraction tasks such as named entity recognition and relation (event) extraction depend heavily on manually-annotated corpora. Although several corpora have been developed for specialized biomedical domains [64, 282, 149, 5], there is still a need for a corpus that can bridge biological, general scientific, environmental, and clinical scientific sub-languages. We aimed to fill this gap by developing both a specialized ontology and a carefully curated corpus.

**Semantic ambiguity** Semantic ambiguity is pervasive in biomedical prose. When semantic meanings of a word or phrase are clearly separated (as of *bank* in *the East bank of Danube* versus *the Deutsche Bank*), an automated sense disambiguation can be implemented using machine learning tools. Unfortunately, alternative meanings are not always clearly separated in biomedical texts. The problem is not that a phrase can refer to several distinct entities in the real world in different contexts, but that the scientist who wrote a given article may not have bothered to separate competing close meanings in her mind. For example, in molecular biology and genetics, in some contexts, a named entity may refer to a *gene* or a *protein* nearly with equal probability, as in "a mutant hemoglobin $\alpha_2$" can refer to either a gene or a protein. If the author meant *gene-or-protein A* and we force an annotator to choose either interpretation *gene A* or *protein A*, the resulting annotation would be of limited utility because the choice between *gene* and *protein* in this case is random. Ideally, a specialized

ontology of text entities would allow an annotator to choose the proper level of granularity of annotation (*gene-or-protein*, in our example) and minimize the need for forced, random decisions.

**Objective** We have constructed a new ontology specifically for annotating text entities, trying to minimize unwarranted arbitrary assignments of semantic labels by annotators. Using this ontology, we annotated a large biomedical corpus to enable a broad spectrum of natural language processing and biomedical machine learning tasks. Our corpus differs from previous efforts in several significant aspects. The Named Entity Recognition Ontology (NERO) and our annotated corpus aim to encompass all entity types that might occur in biomedical literature. In addition to Named Entities, the ontology captures *events* representing a spectrum of relationships between biomedical concepts.

## 4.2 Results

In total, we annotated 35,865 sentences. These sentences encapsulated 190,679 Named Entities. The frequencies of all diverse entity types in our corpus are show in Figure 4.1.

**By Frequency** The most frequent entity type was *GeneOrProtein* accounting for 14.7% of all Named Entities in the corpus (Figure 4.1). The second most populous category was *Process* with 9.0% tagged. *Process* has six sub-concepts and almost half of *Process* instances (49.7%) were annotated as more specific sub-concepts; the *BiologicalProcess* and the *MolecularProcess* were the $5^{th}$ and $7^{th}$ most frequent entity types (Figure 4.1). The frequencies of Entity type frequencies decrease rapidly, with the least frequent types in our corpus being *Journal*, *Unit* and *Citation*, see Figure 4.1.

Our ontology (Figure 4.2) and annotations were designed to capture named entities as-

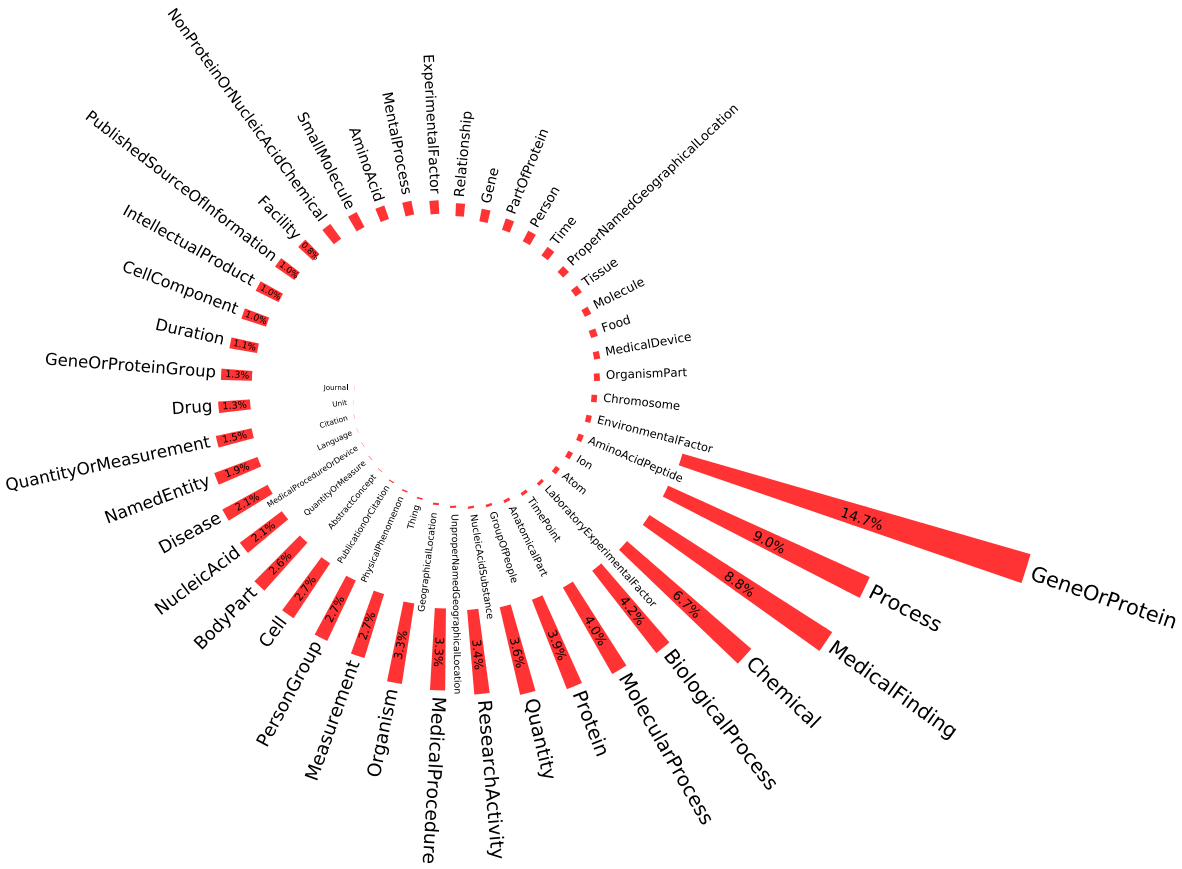


**Figure 4.1: Frequencies of annotated Named Entities in our corpus.**

sociated with research activities and facilities; these types of entities can be important for encoding methods used in scientific experiments or patient treatment. Semantic classes of *ResearchActivity* and *MedicalProcedures* turned out to be the $9^{th}$ and the $10^{th}$ most frequent, respectively. Other top concepts related to research included *Measurement*, *IntellectualProducts*, *PublishedSourceOfInformation*, *Facility*, and *MentalProcess*.

We tried fitting the rank-ordered frequency distribution of annotated Named Entities with a Discrete Generalized Beta Distribution (DGBD), see Figure 4.3. The result showed a deviation from the Zipf's law [327]: The tail of the observed distribution was not heavy enough to match Zipf's distribution, most likely due to relatively small number of classes in our ontology.[159] In other words, we expect that frequencies of semantic classes in a very large corpus annotated with classes from a hypothetical perfect Named Entity Ontology would follow a Zipfean (discrete Pareto) distribution of named entity classes.

**By Branches** In the NERO ontology, under the *NamedEntity* cluster (Figure 4.2), in addition to the almost two dozen more sparsely used branches (such as *ExperimentalFactor* and *GeographicalLocation*), there are three branches heavily represented in our corpus: *AnatomicalPart*, *Chemical*, and *Process*. Slightly more than half (51.6%) of all entities are from these three classes, with 26.6% of all entities originating from *Process* alone.

Ambiguity levels varied broadly across Named Entities captured in our corpus. For example, in class *AnatomicalPart*, almost all (99.3%) were annotated at the most specific levels, with the majority of entities belonging to *BodyPart*, *CellularComponent*, and *Cell*. In contrast, the general (most vague) concept, *Chemical*, turned out to be the most annotated within its cluster, although more specific subclasses, such as *Protein*, *NucleicAcid* and *Drug* were also well represented in the corpus.

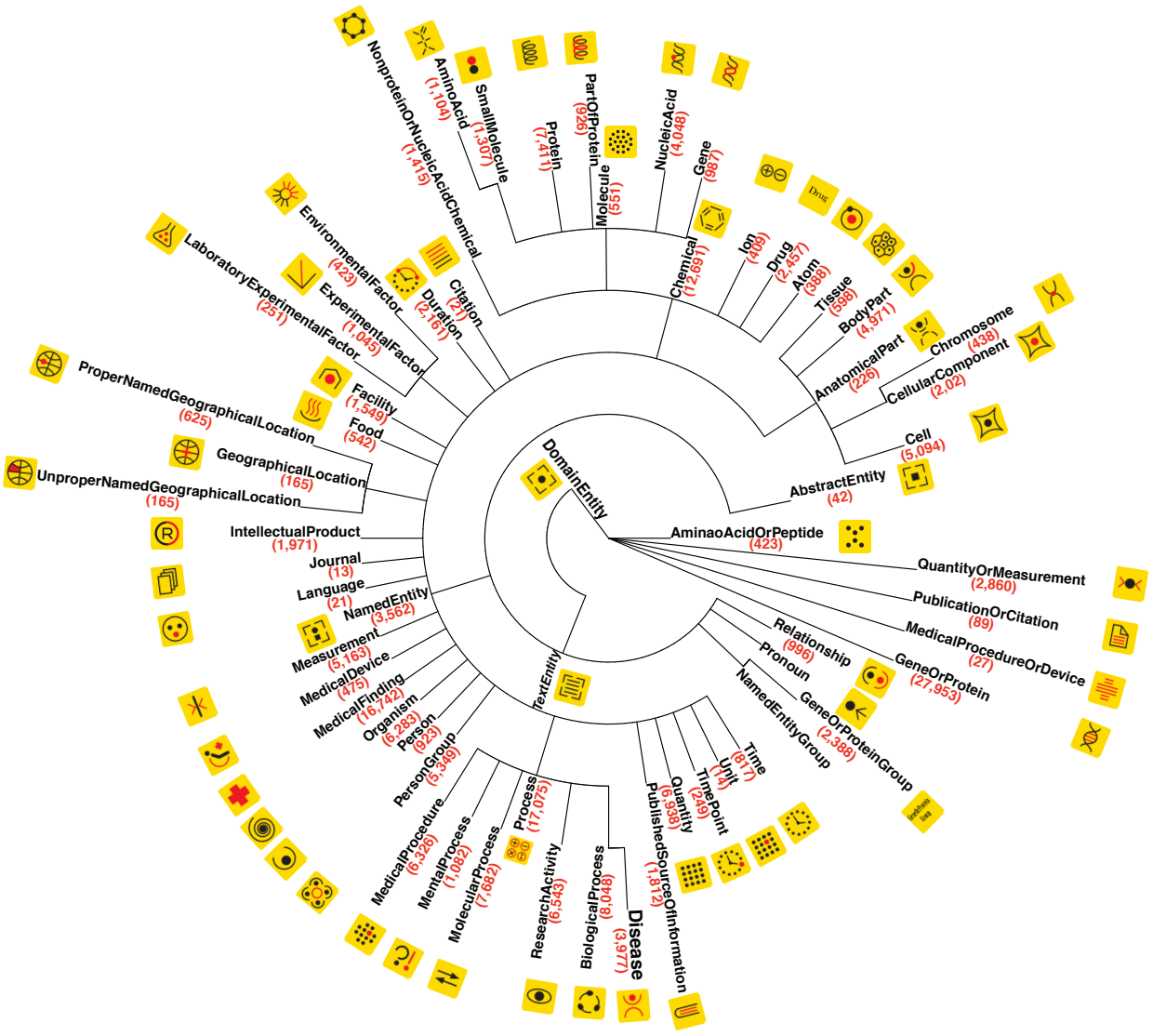


**Figure 4.2: Named Entity Recognition Ontology (NERO).** The Ontology is represented with class names and icons, and annotation frequencies are shown in parentheses next to each Named Entity class.

In the *Process* concept cluster, about a third of all concept instances were annotated at a more general *Process* level and the rest of them were specific concepts, such as *MedicalProcedure*, *MolecularProcess*, *ResearchActivity*, and *BiologicalProcess*.

In addition to these major clusters of concepts, several individual concepts are well represented in the corpus. For example, *MedicalFinding* represents 7.3% of all entities. Other well-represented concepts included *Duration*, *IntellectualProduct*, *Measurement*, *Organism*, *PersonGroup*, *PublishedSource OfInformation*, and *Quantity* (Figure 4.1). In total, about 70.4% of all entities were annotated at the most specific ontology level (Figure 4.2).

There are five concepts in NERO ontology that allow semantic flexibility in order to avoid arbitrary concept assignment. Entities annotated as *AminaoAcidOrPeptide*, *QuantityOrMeasurement*, *PublicationOrCitation MedicalProcedureOrDevice* and *GeneOrProtein* account for 17.8% of all entities. And less than a quarter (23%) of entities representing either genes or proteins were cleanly annotated with class *Gene* or class *Protein*, the rest of them were annotated with class *GeneOrProtein*.

**Actions (Events)** In addition to the 190,679 Named entities, we annotated 43,438 action terms (events connecting two or more entities). The most annotated action term is *bind*, accounting for 28.4% of all actions, Figure 4.4. When we normalize the action terms and combine actions such as *bind*, *binds*, *binding*, the normalized action *bind* accounts for 31.8% of all actions, as shown in Figure 4.4. In addition to the action *bind*, actions indicating entities' attributes are the next most frequent. Other biological relationships are also well represented in this annotation, such as *inhibit*, *activate*, *mediate*, *interact*, *contain*, and *regulate*. The top 30 action categories accounted for 64.4% of all actions annotated with the top 10 action categories accounting for 52.2%.

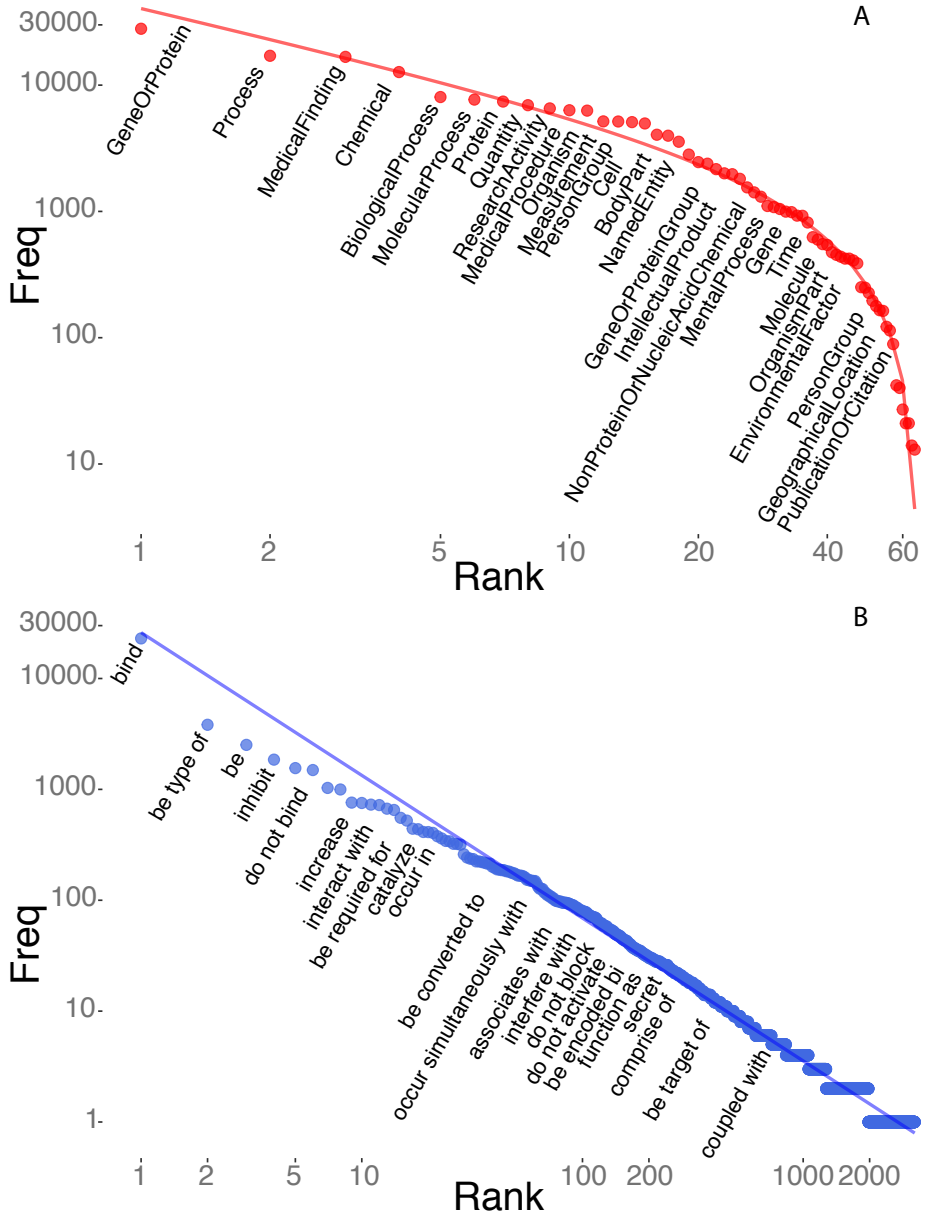


**Figure 4.3: Frequencies of Named Entity and Actions by Rank.** (A) Named Entity term frequencies, curve fitted with Discrete Generalized Beta Distribution (DGBD). (B) Action term frequencies, curve-fitted with Zipf's distribution.

Interestingly, negations of actions were also quite abundant in our annotated corpus. For example, *do not bind* was the $6^{th}$ most frequent normalized action. Other well-represented negations of actions include *do not affect* and *do not inhibit* (Figure 4.4). The rank-ordered distribution of annotated Actions were fitted with a Zipf Distribution, shown in Figure 4.3. The result showed no clear deviation from the Zipf's law [327] as actions types are not limited.

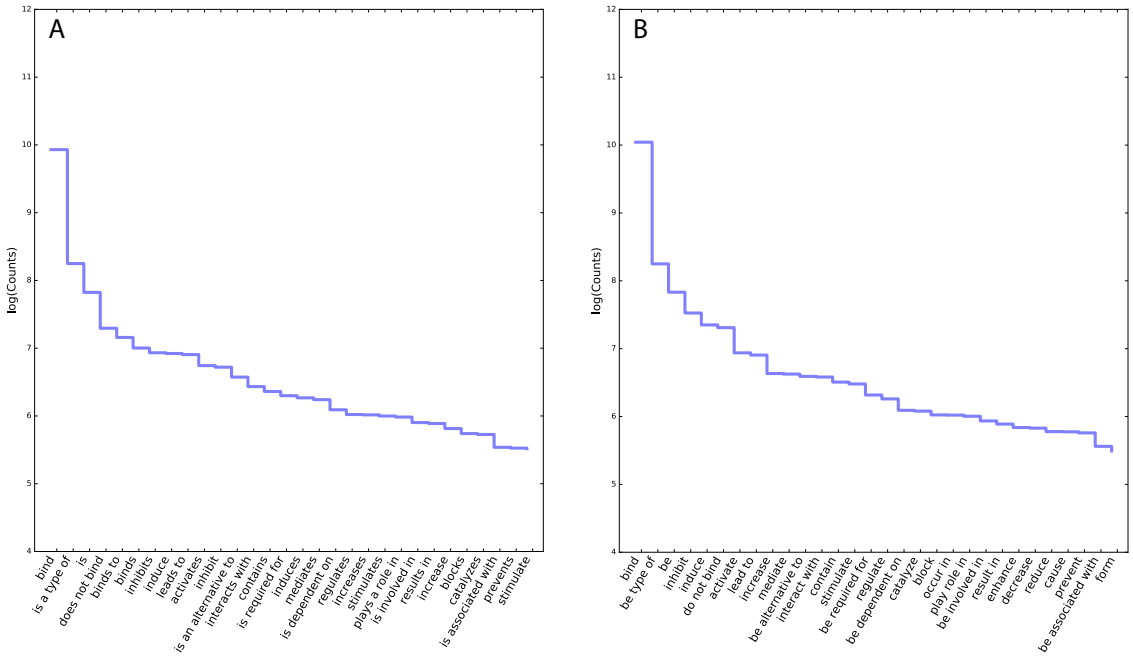


**Figure 4.4: Top 30 Actions in corpus** (A) Without normalization (B) With Normalization

### *4.2.1 Interannotator-Agreement Statistics*

8,650 (24.1%) of the 35,865 sentences were annotated by multiple annotators. Table 4.1 illustrates that our annotators maintained consistently high inter-annotator agreement.

| | Agreement Type | IAA(%) |
|---|---|---|
| Named Entity Span | Exact Match | 86.49 |
| | Relaxed Match | 93.66 |
| Concept Assignment | Exact Match | 86.56 |
| | Parent Match | 87.66 |
| | Superclass Match | 86.72 |
| | Ambiguity Match | 97.58 |

**Table 4.1: Inter-annotator Agreement Statistics.**

### *4.2.2 Named Entity detections: Classifier performance*

We conducted two initial machine learning experiments. Using NERsuite, we conducted 10-fold cross-validation, dividing corpus into training and test subsets. The classification results are presented in Table 4.2. The overall performance is moderate, with 54.9% precision, 37.3% recall and 43.4% $F_1$. The best performance class is GeneOrProtein with baseline results of 67.0% precision, 65.3% recall, and 66.2% $F_1$ score.

We then trained an additional set of classifiers on our corpus data for the top 20 classes. We randomly choose 90% of the sentences to be the training set, and the remaining 10% to be the test set. Then we use this model to tag semantic entities in a fresh set of 141,822 PubMed articles. We obtained the following statistics (Table 4.3).

The resulting precision is fair (51% overall) while recall is lower (42% overall) with an overall F1 score of 46%. The Precision is 68% (PersonGroup) on the high end and 23% on the low end. Recall performance varied significantly, with 61%(GeneOrProtein) as the highest and 9% as the lowest. Overall for F1 score, *GeneOrProtein* entities were associated with the best performance of NER engine, 66.38%.

| | Baseline | | | Baseline-Dict Features | | | Stacking | | | Merging | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **P**(%) | **R**(%) | $F_1$(%) | **P**(%) | **R**(%) | $F_1$(%) | **P**(%) | **R**(%) | $F_1$(%) | **P**(%) | **R**(%) | $F_1$(%) |
| Cell | 62.79 | 56.01 | 59.17 | 62.17 | 55.28 | 58.48 | 62.84 | 56.75 | 59.60 | 60.14 | 53.44 | 56.54 |
| CellComponent | 59.01 | 41.40 | 48.58 | 58.98 | 41.88 | 48.90 | 58.13 | 41.29 | 48.19 | 54.75 | 39.61 | 45.91 |
| GeneOrProtein | 67.00 | 65.35 | 66.16 | 67.05 | 65.81 | 66.42 | 67.02 | 66.04 | 66.52 | 68.33 | 63.52 | 65.83 |
| Organism | 71.72 | 55.14 | 62.32 | 71.35 | 57.00 | 63.33 | 71.03 | 55.70 | 62.40 | 69.73 | 52.58 | 59.92 |
| Disease | 69.72 | 54.75 | 61.28 | 69.21 | 55.11 | 61.29 | 70.23 | 56.93 | 62.83 | 68.63 | 50.72 | 58.26 |
| Drug | 64.13 | 40.40 | 49.43 | 64.88 | 42.95 | 51.59 | 62.19 | 42.51 | 50.35 | 59.60 | 44.18 | 50.64 |
| SmallMolecule | 26.84 | 6.04 | 9.77 | 24.09 | 5.57 | 8.94 | 23.70 | 5.79 | 9.17 | 17.94 | 4.13 | 6.67 |
| BiologicalProcess | 46.03 | 26.64 | 33.71 | 46.08 | 27.23 | 34.19 | 46.19 | 27.24 | 34.23 | 45.71 | 21.07 | 28.81 |
| MolecularProcess | 40.67 | 26.01 | 31.70 | 40.64 | 25.78 | 31.52 | 40.92 | 25.90 | 31.70 | 41.19 | 18.80 | 25.79 |
| Gene | 49.35 | 16.490 | 24.32 | 47.59 | 16.17 | 23.6 | 49.94 | 16.76 | 24.62 | 28.81 | 11.73 | 16.36 |
| Protein | 44.17 | 25.72 | 32.49 | 44.91 | 26.22 | 33.09 | 45.10 | 26.48 | 33.35 | 37.25 | 25.27 | 30.10 |
| BodyPart | 64.62 | 49.02 | 55.72 | 65.05 | 50.30 | 56.70 | 65.23 | 50.13 | 56.67 | 66.75 | 42.86 | 52.18 |
| AminoAcid | 47.53 | 22.37 | 30.20 | 48.88 | 23.24 | 31.29 | 45.15 | 21.29 | 28.72 | 48.10 | 21.84 | 29.75 |
| **overall** | 54.89 | 37.33 | 43.45 | 54.68 | 37.89 | 43.80 | 54.44 | 37.91 | 43.72 | 51.30 | 34.60 | 40.52 |

**Table 4.2: Experimental results for NER evaluated on 10% of the corpus.**

| Entity | P | R | F1 | # Papers contain entity |
|---|---|---|---|---|
| BiologicalProcess | 0.37 | 0.29 | 0.32 | 135,589 |
| BodyPart | 0.64 | 0.51 | 0.57 | 115,614 |
| Cell | 0.62 | 0.53 | 0.57 | 110,843 |
| Chemical | 0.51 | 0.39 | 0.44 | 123,115 |
| Disease | 0.59 | 0.53 | 0.56 | 94,871 |
| Drug | 0.61 | 0.47 | 0.53 | 62,444 |
| GeneOrProtein | 0.64 | 0.61 | 0.62 | 126,881 |
| Measurement | 0.44 | 0.23 | 0.3 | 107,732 |
| MedicalFinding | 0.49 | 0.44 | 0.46 | 138,191 |
| MedicalProcedure | 0.57 | 0.51 | 0.54 | 122,666 |
| MolecularProcess | 0.39 | 0.27 | 0.31 | 128,828 |
| NamedEntity | 0.23 | 0.09 | 0.13 | 131,084 |
| NucleicAcid | 0.46 | 0.29 | 0.36 | 94,819 |
| Organism | 0.63 | 0.49 | 0.55 | 117,449 |
| PersonGroup | 0.68 | 0.66 | 0.67 | 128,380 |
| Process | 0.38 | 0.37 | 0.37 | 140,806 |
| Protein | 0.42 | 0.26 | 0.32 | 113,335 |
| Quantity | 0.38 | 0.33 | 0.36 | 136,014 |
| QuantityOrMeasurement. | 0.25 | 0.11 | 0.16 | 117,891 |
| ResearchActivity | 0.61 | 0.49 | 0.54 | 137,301 |
| Totals | 0.51 | 0.42 | 0.46 | 141,822 |

**Table 4.3: Experimental results for NER evaluated for the top 20 annotated.**

## 4.3 Methods

### *4.3.1 Nero Ontology*

The topic area of the Named Entity Recognition Ontology (NERO) is the lexical representation of entities, rather than the entities themselves. For example, we want vocabulary to represent the set of protein names found in a text, rather than the protein that information content represents. Thus, the main aim of NERO is to enable text annotators or text annotation tools to mark up the a text's lexical content as to the nature of that lexical entity.

For example, in the sentence:

> Activation of NF-$\kappa$B2 and RelB was found in 53.7 and 49.2 % of the 121 ER+ tumours analyzed, with similar levels to ER-breast tumours analysed in parallel for comparisons. [237]

Here, NF-$\kappa$B2 and RelB can be either a gene or protein.

In gene and protein naming conventions, italics is used for genes and mRNAs and normal text is used for proteins. More specifically, human genes and proteins are all capitalized. Mice and rat gene symbols have the first letter capitalized, while protein symbols are all upper-case. In contrast, for flies, both gene and protein symbols can begin with an upper-case letter. However, researchers do not follow these naming conventions strictly and often use the same symbol to represent both a gene encoded for a protein and the protein itself.

An annotator, if forced to commit to either a gene or a protein, risks mis-annotating. Enabling an annotator to commit less strongly by annotating these lexical entities as 'gene or protein named entity' avoids such a risk, but still allows annotations to be made and

queries posed and answered.

In NERO, we would like to cover all entity types that might occur in biomedical articles. We start, however, with entities around molecules and their interactions within a cell, their link to disease, and the machinery or tests used to investigate these entities.

Thus, the basic competencies for NERO are:

1. Provide vocabulary for annotating the entities covered in the scope outlined above.

2. Provide abstractions of the lexical items such that annotators can commit to an annotation with an appropriate confidence level.

3. To include knowledge about which biological or domain entity a given text entity represents.

NERO is authored in OWL DL using the **protege** 4 authoring environment.NERO may be downloaded with a license. NERO is a simple ontology; it is not axiomatized highly; it only requires a simple taxonomy to fulfill the competencies above. The main features of NERO can be seen in Figure 4.2. We use a naming convention in which all class labels end with the suffix 'entity.' Labels also capitalize the initial letter. NERO covers text entities and hence *DomainEntity*–and all semantic ambiguous classes–sits around the NERO's root. The basic division thereafter is into *TextEntity* and *AbstractEntity*, where *TextEntity* further split into *NamedEntity*, *NamedEntityGroup*, *Relationship* and *Pronoun*. The pronouns amount to a set of commonly occurring English pronouns.

After *NamedEntity*, the hierarchy essentially reflects that which may be seen in many descriptions of biological entities, rather than in the lexical representation of those entities. NERO differs in cases such as *'GeneOrProtein'*, which subsumes both *Gene* and *Protein* using the following axiom: *EquivalentTo: 'Gene' or 'Protein'*. There are no biological entities that are either a gene or a protein, but there are lexical entities that are either a gene or a protein. NERO uses this pattern to express ambiguity between various text entities.

Classes in NERO represent information and not the actual biological entities that the information describes. It is, therefore, straight-forward to link between the lexical or information [informational? EG] entity and the biological entity through a relationship such as *'is about'*. So the NERO class *Protein 'is about' some 'protein'* in an ontology such as the Protein Ontology([200]).

### *4.3.2 Annotation*

**Data Sources and Preperation** The annotation on the corpus was performed by 10 Ph.D.-level annotators with deep experience in biomedical text annotation or biomedical research. Each annotator was first trained on a practice set of 200-300 sentences before moving on to the 'production' annotation stage. The entire corpus consists of 35,865 sentences from 8,080 MEDLINE-referenced articles or abstracts. The sentences are selected for annotators randomly.

**Annotation Guidelines** The guidelines for annotation practice have been developed by early annotators and further discussed and finalized. Any changes to the guidelines were discussed thoroughly, and annotators were informed of those changes made in each version. We aimed to annotate Named Entities relevant to biomedicine as represented in the NERO Ontology. We intended to capture Named Entities at the most specific level on the ontological

tree. See Appendix for the complete guidelines.

**Annotation Process and Interface** In order to facilitate the annotation process, we developed a web-based annotation tool. First, annotators read the sentences. Below each sentence is a group of Named Entity classes represented in graphic icons (Figure 4.5). Annotators then assign a class for each relevant Named Entity by dragging the icon to the Semantic class. To ensure the class consistency, the Semantic class can only be filled using the icon; annotators are not able to enter the Named Entity class manually (it is greyed out). The annotation tool also allows annotators to annotate modifiers for the Named Entities as well. When two Named Entities interact, annotators were able to annotate the action terms.

After the initial annotation, a second annotator reviews the sentences and annotations. Disagreements were discussed (and occasionally resolved) with a third annotator for any remaining discrepancies. To explain this process, we used the following sentence as an example:

> Cyclosporin A (CsA) abrogated the binding of Sp1 and NFAT1 to the p21WAF1/CIP1 promoter in high Ca2+-treated NHK cells, restoring the binding of KLF16, as assayed by a chromatin immunoprecipitation assay. [248]

The annotation results are shown in Figure 4.6

**Inter-Annotator Agreement** In order to assess the reliability of our annotations, a portion of the corpus was assigned to multiple annotators. We evaluated the process for the following annotation subtasks:

- *Exact* span matches, where two annotators identified exact the same Named Entity text spans.
- *Relaxed* span matches, where Named Entity text spans from two annotators overlap.

- *Exact* concept matches, where within agreed text span, annotators assigned exact same concept class.
- *Parent* concept matches, where the concept class assigned by one annotator is the parent class of the one by the other annotator.
- *Superclass* concept matches, where the two concept classes assigned belong to the same superclass
- *Ambiguity* concept matches, where one annotator assigned a semantic ambiguous class which includes the concept assigned by the other annotator.

Due to the difficulty in defining the size of negative annotations, instead of $\kappa$ statistic, we reported inter-annotator agreement(IAA) using positive specific agreement or F-measure following the formula from [128].

### *4.3.3 Semantic Classes*

***AbstractConcept*** A named entity that can have many meanings. This class is a SUPERCLASS of other classes below. It can be used to define the boundaries of a named entity when more detailed class assignment is difficult. All proper noun phrases that are not better matched as one of the other classes are to be assigned an abstract concept. In those cases where a phrase can be assigned to more than one class, the abstract concept is to be used instead (e.g. Washington could be a person or a location, or Cell could be the name of the Journal or refer to a biological cell – in both cases, *AbstractConcept* should be used.)

***Time*** A period or time point. A calendar time description that includes the year, decade, or century. Other phrases that describe a duration or time related concept are also in this class.

- Polychlorinated biphenyls (PCBs) were measured in the air and water over the Hudson River Estuary during six intensive field campaigns from December 1999 to April 2001.

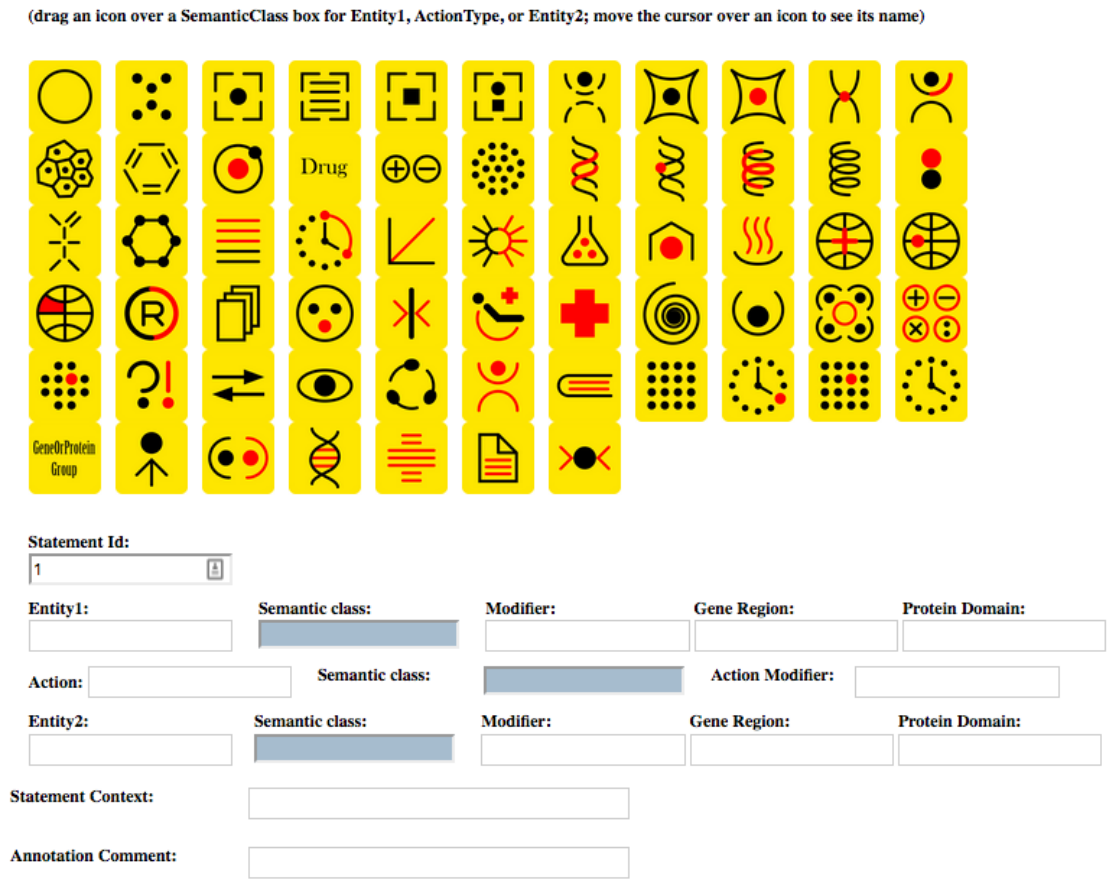


**Figure 4.5: Web annotation tools**

- The international skeletal society meeting, Budapest 2007: special scientific and radiological focus program, tuesday, 9 october 2007.
- Alois Alzheimer (1864 – 1915) presented the first case of a patient with symptoms of a disease that later would be called Alzheimer's disease.

***GeographicalLocation*** A proper name of a geographical location.

- Both poles of Mars are hidden beneath caps of layered ice.
- Polychlorinated biphenyls (PCBs) were measured in the air and water over the Hudson River Estuary during six intensive field campaigns from December 1999 to April 2001.

| id | mod1 | g1 | p1 | entity1 | sem1 | action modifier | action | sem action | mod2 | g2 | p2 | entity2 | sem2 | context | user | editor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Statement[1] | | | | Sp1 | GP | | bind | | | promoter | | p21WAF1/CIP1 | GP | in high Ca2+-treated NHK cells | res | cle |
| Statement[2] | | | | NFAT1 | GP | | bind | | | promoter | | p21WAF1/CIP1 | GP | in high Ca2+-treated NHK cells | res | cle |
| Statement[3] | | | | CsA | Chemical | | abrogate | | | | | statement 1 | Process | | res | cle |
| Statement[4] | | | | CsA | Chemical | | is | | | | | Cyclosporin A | Chemical | | res | cle |
| Statement[5] | | | | CsA | Chemical | | abrogate | | | | | statement 2 | Process | | res | cle |
| Statement[6] | | | | statement 3 | Process | | restore | | binding of | | | KLF16 | GP | | res | cle |
| Statement[7] | | | | statement 5 | Process | | restore | | binding of | | | KLF16 | GP | | res | cle |
| Statement[1] | | | | Sp1 | GP | | bind | | | promoter | | p21WAF1/CIP1 | GP | in high Ca2+-treated NHK cells | sub | cle |
| Statement[2] | | | | NFAT1 | GP | | bind | | | promoter | | p21WAF1/CIP1 | GP | in high Ca2+-treated NHK cells | sub | cle |
| Statement[3] | | | | CsA | Chemical | | abrogates | | | | | statement 1 | Process | | sub | cle |
| Statement[4] | | | | CsA | Chemical | | abrogates | | | | | statement 2 | Process | | sub | cle |
| Statement[5] | | | | CsA | Chemical | | is | | | | | Cyclosporin A | Chemical | | sub | cle |
| Statement[6] | | | | statement 3 | Process | | restore | | binding of | | | KLF16 | GP | | sub | cle |
| Statement[7] | | | | statement 4 | Process | | restore | | binding of | | | KLF16 | GP | | sub | cle |

**Figure 4.6: Web Annotation Result Example**

- Hospitalizations of patients with acute rheumatic fever were significantly more common in the Northeast and less common in the South.

- **Nested geographic** Thirty-eight patients with chronic heart failure, age 57+/-2 years, New York Heart Association classification II-III, were assigned to either a high intensity training group (n=15, age 53+/-2 years) exercised at 60% of sustained maximal inspiratory pressure, or a low intensity training group (n=23, age 59+/-2 years), exercised at 15% of sustained maximal inspiratory pressure, three times per week for 10 weeks.

***UnproperNamedGeographicalLocation*** A geographical location that is not a proper name.

- Patterns of bacterial diversity across a range of Antarctic terrestrial habitats.

- The net H(+) production associated with Al and Fe transformations was 252 and 1meqm(-2)yr(-1) (on the lake area basis), respectively, reflecting fluxes of ionic, organic,

and particulate forms into and out of the lake and the pH gradient between the inlet and outlet.

- Both poles of Mars are hidden beneath caps of layered ice.
- Effect of restricted suckling on milk yield, milk composition and udder health in cows and behaviour and weight gain in calves, in dual-purpose cattle in the tropics.

***PersonGroup*** A proper name of an association of individuals, including companies, clubs, political organizations, government branches, or other entities, as well as groups by character, of people that share certain characteristics such as profession, gender, nationality, or disease.

- Thirty-eight patients with chronic heart failure, age 57+/-2 years, New York Heart Association classification II-III, were assigned to either a high-intensity training group (n=15, age 53+/-2 years) exercised at 60% of sustained maximal inspiratory pressure, or a low-intensity training group (n=23, age 59+/-2 years), exercised at 15% of sustained maximal inspiratory pressure, three times per week for 10 weeks.
- Residents of this valley are predominantly nonsmoking members of the Church of Jesus Christ of Latter-day Saints (Mormons).
- Epigenomics and disease, tenth anniversary winter meeting of the UK Molecular Epidemiology Group (MEG), The Royal Statistical Society, London, UK, 8th December 2006.
- lawyers, physicians, journalists
- diabetics, hurricane victims
- Americans, Russians, Spaniards
- educated people, people of good will

***Person*** A proper name of an individual person.

- Arsenic speciation of two specimens of Napoleon's hair.
- Ten registered Democrats and ten registered Republicans were scanned in an event-related functional MRI paradigm while viewing pictures of the faces of George Bush, John Kerry, and Ralph Nader during the 2004 United States presidential campaign.
- Alois Alzheimer (1864-1915) presented the first case of a patient with symptoms of a disease that later would be called Alzheimer's disease.

***Organism*** An organism, including plant, alga, fungus, virus, bacterium, archaeon, and animal. Including developmental and post-mortem stages. Also covers humans as organisms.

- Arsenic speciation of two specimens of Napoleon's hair.
- Deficiency in recapitulation of stage-specific embryonic gene transcription in two-cell stage cloned mouse embryos.
- Interference competition between introduced black rats and endemic Galápagos rice rats.
- Effect of restricted suckling on milk yield, milk composition, and udder health in cows and behaviour and weight gain in calves, in dual-purpose cattle in the tropics.

Does not cover humans as individual persons (which would fall under Person.)

***AnatomicalPart*** A multi-cellular organization or location of an organ. It includes body parts, body location, body regions, body fluids, organs, organ components, tissue, anatomical structure.

-

- Ten registered Democrats and ten registered Republicans were scanned in an event-related functional MRI paradigm while viewing pictures of the faces of George Bush, John Kerry, and Ralph Nader during the 2004 United States presidential campaign.

- Two months later, during follow-up, a chest X-ray and computed tomography documented a coin lesion of the upper left lung, confirmed by positron emission tomography.

- Non-suicidal self-injury is the intentional destruction of body tissue without suicidal intent and for purposes not socially sanctioned.

Does not include organisms living within organisms.

***Cell*** A cell or cell line that is not an organism.

- Children with autoimmune disease and CNS injury also exhibited abnormal T-cell responses against multiple cow-milk proteins.

- Neonatal and adult microglia cross-present exogenous antigens.

- The effects of the LABAs salmeterol and formoterol on the synthesis of soluble interleukin-8 (IL-8), granulocyte-macrophage colony-stimulating factor (GM-CSF), and vascular endothelial growth factor (VEGF) in the human airway epithelial cell line A549 was investigated in vitro.

***CellularComponent*** A sub-cellular structure that is neither a gene nor a protein nor a nucleic acid structure.

- Putative, full-unit length begomoviral DNA multimers were digested with Nco I and cloned into the plasmid vector pGEM7Zf+.

- Vinculin links integrin receptors to the actin cytoskeleton by binding to talin.

- Caveolae are extremely stable elements of PECs and can be excluded from their cell membrane only in response to the dramatic cell reconstruction observed in FSGS and LGN.

- Changes in cell morphology and cytoskeletal organization are induced by human mitotic checkpoint gene, Bub1.

***GeneOrProtein*** A gene or protein name, including peptides, but excluding partial sequences. This class also includes secondary structures like alpha sheets and beta coils.

- Children with autoimmune disease and CNS injury also exhibited abnormal T-cell responses against multiple cow-milk proteins.

- The effects of the LABAs salmeterol and formoterol on the synthesis of soluble interleukin-8 (IL-8), granulocyte-macrophage colony-stimulating factor (GM-CSF), and vascular endothelial growth factor (VEGF) in the human airway epithelial cell line A549 was investigated in vitro.

- Liposomes incorporating a Plasmodium amino acid sequence target heparan sulfate binding sites in liver.

***GeneOrProteinGroup*** A group of proteins or gene clusters

- Gene expression of CYP3A4 , ABC-transporters (MDR1 and MRP1-MRP5), and hPXR in three different human colon carcinoma cell lines.

- Genome sequence analysis of Streptomyces ambofaciens ATCC23877 has revealed numerous secondary metabolite biosynthetic gene clusters, including a giant type I modular polyketide synthase (PKS) gene cluster, which is composed of 25 genes (nine of which encode PKSs) and spans almost 150 kb, making it one of the largest polyketide biosynthetic gene clusters described to date.

***AminoAcid*** An amino acid name or sequence of amino acids. This class also includes small peptides that that map to a part of a protein-coding gene or single amino acids.

- Liposomes incorporating a Plasmodium amino acid sequence target heparan sulfate binding sites in liver.
- Cleaving Ala(444)-Ala(445) released mini-plasmin with secondary activity to hydrolyze fibrin.
- Mass spectrometry analysis of the Ebola virus soluble glycoprotein sGP identified a rare post-translation modification, C-mannosylation, which was found on tryptophan (W) 288.

Peptides that are stand-alone would fall under Gene-or-Protein.

***NucleicAcid*** A chemical structure that is based on nucleic acids. It includes nucleoside, nucleotide, RNA, DNA, sites in a sequence, and artificially constructed sequences such as vectors or plasmids.

- Putative, full-unit length begomoviral DNA multimers were digested with Nco I and cloned into the plasmid vector pGEM7Zf+.
- The deletion occurred at the consensus cleavage site (3'-A—TTTT-5') without target site duplication.
- The very long telomeres in Sorex granarius (Soricidae, Eulipothyphla) contain ribosomal DNA.

Does not include chromosomes or genes (the latter would fall under Gene-or-Protein).

***Chromosome*** A chromosome, chromosome region, chromosome part, or chromosome position. It does not include chromosome positions that can be considered measure in units (e.g., 300 bp).

- The very long telomeres in Sorex granarius (Soricidae, Eulipothyphla) contain ribosomal DNA.
- No evidence of linkage between 7q33-36 locus (OTSC2) and otosclerosis in seven British Caucasian pedigrees.
- Failure to confirm allelic and haplotypic association between markers at the chromosome 6p22.3 dystrobrevin-binding protein 1 (DTNBP1) locus and schizophrenia.

***NonNucleicAcidNonProteinChemical*** A chemical structure or a material that is not a gene, a protein, an amino acid, a chromosome, or based on nucleic acids. It includes chemical elements, ions, isotopes, organophosphorus compounds, carbohydrates, lipids, pharmacological substances, and drugs. Drugs are recorded under this category, even when the drug's composition substances are unknown.

- Polychlorinated biphenyls (PCBs) were measured in the air and water over the Hudson River Estuary during six intensive field campaigns from December 1999 to April 2001.
- The net H(+) production associated with Al and Fe transformations was 252 and 1meqm(-2)yr(-1) (on the lake area basis), respectively, reflecting fluxes of ionic, organic, and particulate forms into and out of the lake and the pH gradient between the inlet and outlet.
- Neonatal and adult microglia cross-present exogenous antigens.

Does not include food.

***Food*** Food or drink that is not a simple substance (e.g., salt, water) or an organism name (e.g., wheat, pig, rice).

- The 2005 White House Conference on Aging: a new day for White House conferences on aging and food for the future.
- When comparing highest versus lowest levels of intake in multivariable adjusted models, positive associations were observed for several beef / lamb and individual animal protein items, including beef / lamb as a main dish (OR = 2.2, 95% CI: 1.0-4.5), regular hamburger (OR = 1.7, 95% CI: 1.2-2.4), whole eggs (OR = 1.6, 95% CI: 1.0-2.4), butter (OR = 2.4, 95% CI: 1.6-3.5), and total dairy not including butter (OR = 2.6, 95% CI: 1.8-3.7).
- Digestion rate of legume carbohydrates and glycemic index of legume-based meals.

***EnvironmentalFactor*** Environmental factor

- UV light, radiation ...

***Relationship*** Phrases that express or imply a relationship between objects.

- Mathematical, statistical, or logical relationships: correlation, causation, dependency, equality, progression, significant difference, inverse ...
- Comparisons: similarity, dissimilarity, commonality, increased risk
- Kinships: descendant, ancestor, sibling ...

***Process*** A general, organismal, cellular, or chemical process. This includes processes on the organismal level involving whole tissues or groups of cells such as growth and pathogenesis. It includes processes at the cell level or involving sub-cellular components (e.g., organelles), such as differentiation or apoptosis.

- Caveolae are extremely stable elements of PECs and can be excluded from their cell membrane only in response to the dramatic cell reconstruction observed in FSGS and LGN.

- Thyroid hormone receptor-beta (TRbeta1) impairs cell proliferation by the transcriptional inhibition of cyclins D1, E, and A2.

- The irreversible nature of mitotic entry is due to the activation of mitosis specific kinases such as cdk1/cyclin B.

- Since wee1 keeps cdk1/cyclin B inactive during the S and G(2) phases, its activity must be down-regulated for mitotic progression to occur.

***MolecularProcess*** An activity or event at the chemical or molecular level, including macromolecules like genes or proteins.

- Arsenic speciation of two specimens of Napoleon's hair.

- Deficiency in recapitulation of stage-specific embryonic gene transcription in two-cell stage cloned mouse embryos.

- Liposomes incorporating a Plasmodium amino acid sequence target heparan sulfate binding sites in the liver.

- Digestion rate of legume carbohydrates and glycemic index of legume-based meals.

- Thyroid hormone receptor-beta (TRbeta1) impairs cell proliferation by the transcriptional inhibition of cyclins D1, E, and A2.

- Differential intracellular distribution of DNA complexed with polyethylenimine (PEI) and PEI-polyarginine PTD influences exogenous gene expression within live COS-7 cells.

- The net H(+) production associated with Al and Fe transformations was 252 and 1meqm(-2)yr(-1) (on the lake area basis), respectively, reflecting fluxes of ionic, organic, and particulate forms into and out of the lake and the pH gradient between the inlet and outlet.

***BiologicalProcess*** An interaction at the level of cellular components, cells, organs, organisms, or populations.

- Interactions of immune cells with bacterial cells, cell differentiation, cell death, apoptosis ...
- Hormonal regulation, organ formation and growth, blood pressure regulation, immune response ...
- Digestion, circulation, breathing ...

***MedicalFinding*** Processes that can be considered a specific Medical-finding are to be covered here. An objectively measured sign or symptom (patient-reported problem), or a medical description or observation or finding related to the state of an organism, including sign, symptom, laboratory or test result, syndrome, disease, neoplastic process, mental dysfunction, behavioral dysfunction, or medical finding that is not a measure in units.

- Confirmatory factor analysis of the Epworth Sleepiness Scale (ESS) in patients with obstructive sleep apnea.
- Alois Alzheimer (1864-1915) presented the first case of a patient with symptoms of a disease that later would be called Alzheimer's disease.
- Additionally, the high cholesterol levels found in atherosclerosis could modulate host immunity.
- Thirty-eight patients with chronic heart failure, age 57+/-2 years, New York Heart Association classification II-III, were assigned to either a high-intensity training group

(n=15, age 53+/-2 years) exercised at 60% of sustained maximal inspiratory pressure, or a low-intensity training group (n=23, age 59+/-2 years), exercised at 15% of sustained maximal inspiratory pressure, three times per week for 10 weeks.

Does not include cellular, sub-cellular, molecular or chemical processes (e.g., apoptosis, glycemic index).

***MedicalProcedureOrDevice*** A laboratory, therapeutic, diagnostic procedure or method.

- Malignant hyperthermia as a complication of general anesthesia in the clinic of maxillofacial surgery.
- Conventional X-ray exposures in a-p and axial projections and an MRI investigation are considered standard parts of the surgical planning, and a CT examination is also performed when bony defects are present.
- Two months later, during follow-up, a chest X-ray and computed tomography documented a coin lesion of the upper left lung, confirmed by positron emission tomography.

A human-made device, including mechanical, electric, or electronic devices.

- Tomorrow's stethoscope: The hand-held ultrasound device?
- The effect of seat belt use on the cervical electromyogram response to whiplash-type impacts.
- Evaluation of a digitally integrated, accelerometer-based activity monitor for the measurement of activity in cats.
- CT

Does not include buildings or other construction or construction parts (e.g., a room), which fall under the class Facility.

***QuantityOrMeasure*** A numeric value with measuring units, or a phrase expressing a concept of quantity, such as score, dose, rate, size, length, weight, and related terms.

- Thirty-eight patients with chronic heart failure, age 57+/-2 years, New York Heart Association classification II-III, were assigned to either a high-intensity training group (n=15, age 53+/-2 years) exercised at 60% of sustained maximal inspiratory pressure, or a low-intensity training group (n=23, age 59+/-2 years), exercised at 15% of sustained maximal inspiratory pressure, three times per week for 10 weeks.
- High fever, shooting pain, inflammation in the throat, elevated blood sugar.

***Facility*** A construction or part of a construction including buildings, bridges, towers, and other man-made edifices.

- The 2005 White House Conference on Aging: A new day for White House conferences on aging and food for the future.
- Effect of hospital volume on outcome of pancreaticoduodenectomy in Italy.
- The huge garbage dump site near the Hsin-Hai Bridge is likely the source of heavy metal pollution.

Note that whole cities fall under *Geographicallocation* instead.

***Journal*** This refers not to an individual copy of the journal, but to the journal as a regularly published source of information. For an individual copy, or article in such a copy, see Publication.

- Cell, PLoS Biology, Bioinformatics, Time, People.

If the context makes it not clear that the word relates to a journal name, the entity will be classified as abstract concept instead.

***Publication*** A paper, manuscript, video, book, diary, note, message, report, letter, journal, etc.

***Language*** Natural and artificial languages, such as English, Spanish, Hebrew, Turkish, Swahili, Fortran, LISP, C++.

***IntellectualProduct*** A patent, idea, concept, hypothesis. The outcome of a mental process. This is not limited to something that might obtain IP-protection, but may include theories, algorithms, conclusions, and the like.

***MentalProcess*** Memory, emotions, thoughts, learning, cognition. Differs from Intellectual-product in that the focus is on the process of thinking or feeling, not on the result of this process.

***ResearchActivity*** Investigation, measurement, validation, MMPI study, running gel, sequencing. Activities that are executed in the process of conducting research. This includes large-scale operations such as clinical trials, as well as individual lab activities such as sequencing. Used instead of the more general *Process*, if it is clear that the process is a research activity. The more specific *MedicalProcedureOrDevice* is applied if it is clear that the research activity is conducted in a medical context. *MentalProcess* is applied if the activity is a mental process instead.

- Confirmatory factor analysis of the Epworth Sleepiness Scale (ESS) in patients with obstructive sleep apnea.

## 4.4 Discussion

**Summary** Named entity recognition, a central task of nearly any natural language processing system, depends heavily on the size, quality, and consistency of an annotated corpus

for training and benchmarking.

Research articles and free-text clinical notes represent a major source of biomedical information. They also represent a formidable challenge for machine reading. In the ideal world, well-structured data, such as diagnostic conclusions and procedures encoded with codes from International Disease Classification, would be seamlessly merged with data extracted from clinical notes and images. To arrive at this state of automated data and text processing, we need to harvest human expertise via annotating high quality corpus.

In order to increase annotation quality, we designed a specialized ontology (NERO). The NERO ontology aims at covering and disambiguating all entity types that might occur in biomedical literature, including genes, proteins, and their interactions, processes, chemicals, research entities, publication entities, and lexical representations of semantic ambiguous entities.

**Semantically Ambiguous Classes** One main motivation for this project was to preserve the semantic ambiguity introduced by the authors to prevent machine-learning applications from learning arbitrary judgments introduced by annotators. This type of semantic ambiguity is prevalent in our corpus–particularly ambiguity between genes and proteins. Not only did *GeneOrProtein* represent the largest group of Named entities in our corpus, its classifier performed best in the two initial machine learning experiments. We believe this type of ambiguity could only be more likely in the clinical notes and other medical texts, as the time constraints to complete those are more stringent.

**Actions** The discovery of etiological links between genetic variants and environmental factors to multiple diseases requires a clear understanding of the relationships between genetic

variation, diseases, patient populations, and treatments.

This project is uniquely positioned to facilitate this task because not only did NERO ontology cover diverse types of concepts, our annotated corpus included action terms representing the relationships between Named entities. Our action annotations have moved beyond interactions between proteins and genes(*e.g.*, *bind*, *inhibit*, *phosphorylate*, *encode*), into interactions involving genetic variants and environmental factors(*e.g.*, *associated with*, *occur in presence of*, *trigger*, *lack*).

**Limitations** One of the limitation of this study is that even though our NERO ontology aimed to cover all entities of the biomedical research literature, we did not cover all levels of granularity in classifying entities. A second limitation is that, while major concepts are well annotated in our corpus, several concept types were not as well represented in our corpus. This could limit the performances of algorithms developed to classify those specific Named entity types.

## 4.5 Acknowledgments

Robert Stevens designed the Named Entity Recognition Ontology. Andrey Rzhetsky conceived and supervised the project, helped design and implement the analyses, and jointly wrote the text with myself and Robert Stevens. Maolin Li, Fenia Christopoulou, Sophia Ananiadou, Jose Luis Ambite, Sahil Garg, Ulf Hermjakob, Daniel Marcu, Emily Sheng, Aram Galstyan designed and performed classification experiments. Larisa Soldatova, Ross King, and James Evans helped revise and edit the text.

# CHAPTER 5

# CONCLUSION

Personalized and precision medicine is a global challenge that requires detailed understanding of a person's genetic, environmental, and lifestyle differences and could potentially modernize biomedical research [127, 201]. It has the potential to accelerate the development of accurate medical diagnosis, efficient therapeutic interventions and cost-effective preventive care. Such a future requires a more accurate and precise disease classification system that properly reflects advances in our understanding of molecular and environmental factors that contribute to disease etiology.

The advances in genotyping and sequencing and the rise of data-intensive medical research have created an unprecedented opportunity to modernize disease classification system that integrate molecular, environmental, and phenotypic data into medical decision making. For example, Consortia such as eMerge, i2b2, UK Biobank and BioBank Japan [276, 92] started to conduct research in the health-care setting and demonstrated the feasibility of integrating molecular information with medical histories and clinical findings. Ideally these projects will expand in scope, encourage collaborations, and scale further to support well-powered studies that forms the foundation for scientific consensus, benefiting various participants(e.g. patients, clinicians, researchers and general public)

In the near future, these studies will lead us to a disease classification system that goes beyond descriptive signs and symptoms and links directly to a deeper understanding of intrinsic biology, disease trajectories, and treatments. Ultimately this classification system will enable the selection of a subset of patients, based on their common biological basis, who are most likely to benefit from a particular medical treatment, such as a drug, or surgical procedure.With the development of an efficient way to continuously validate and incorporate newly emerging biomedical findings, this new system can become a dynamic tool that truly revolutionize patient care and modernize biomedical research

Unfortunately, current system for classifying diseases is actually inhibiting development

of biomedical research. As I described in Chapter 3, multiple different diseases share common genetic and environmental causes, while diseases closely related in underlying biology are classified as far apart. In addition, within one disease (e.g. cancer, autism, and asthma), there can be many subtypes with distinct causes [315, 308, 260, 105, 312, 160, 203, 121, 47, 204, 13]. Current classification system primarily relies on a patient's observable symptoms[309]. At best, advanced taxonomies will include findings from lab or imaging studies or even results from afflicted tissues and cells. However, none are designed to incorporate or harness rapidly emerging molecular data, socio-environmental factors, or related patient characteristics.

Clinically, the failure to incorporate new and precise biological insights causes delayed adoption of new practice guidelines and missed opportunities of effective treatment for specific subgroups. On the scientific side, premature classification and crude phenotyping translate to lower power to detect true signal and higher cost and longer cycle to generate valid scientific insight. This problem is accentuated by a lack of understanding on the environmental effects on diseases. Blinded by significant environmental confounding and limited in detecting genetic effects, studies with current classifications are more likely to waste resources than to make meaningful findings. Therefore, by developing a more careful phenotyping scheme and a more focused approach before devoting large resources and time, we can increase our chance for success.

Diagnosis is the foundation of medicine. Accurately and precisely defining a patient's condition does not guarantee successful care, but it is unequivocally the place to start. Indeed today many tumor types are already molecularly defined. The International Classification of Diseases for Oncology (ICD-O), for example, attempts to incorporate genomic profiles of cancer samples. The WHO/IARC Classification of Tumours series also sought to integrate data from genome screening and other molecular techniques. However, these are only isolated examples of progress, mostly concentrated in cancer classification.

In Chapter 3 of this dissertation, I presented, to the best of our knowledge, the first

study to classify diseases of multiple biological system based on genetic and environmental common factors driving intrinsic biology. As demonstrated in Chapter 3 and by literature, genetic influences on common diseases are most likely complex; each patient's genetic variants will affect his/her disease onset, trajectory, and response to therapeutic measures, while environmental modulation will play a consequential role in these processes.

Although latest breakthroughs have concentrated in genomic studies due to the rapid adoptions of genotyping and sequencing technology, the future may see new hypotheses and tangible progress in understanding disease risk and progression from studying patient's history of exposure to environmental agents, and psychological, social or behavioral information. In Chapter 2 I described a series of studies that sought to delineate environmental effects that could drive multiple diseases. As demonstrated by smoking [65, 324]and other life style studies[152, 68, 9], environmental factors could be easier to manage for the prevention of multiple diseases

Precision and personalized medicine by definition is rich phenotyping medicine. In Chapter 4, I presented a digital phenotyping study attempting to connect multiple data types and go beyond structured clinical data. It is likely that the increased functionality of EHRs and the improved performance of digital tools will open the door to conduct both large cohort studies on a wide range of diseases and targeted studies with moderate sizes focusing on carefully teased out genetic and environmental effects.

If extensive molecular characterization of individuals becomes a routine medical procedure, done before and after the appearance of any disease, it will allow data collection on both diseased and healthy individuals at a vast scale. Future medicine will depend on the successful creation of a system for acquiring and analyzing information of large numbers of individuals relating the molecular profiles, health histories and therapies based on the precise disease trajectories [153]. With the integration of vast and diverse molecular data in medicine, the future may also see the development of "meta-phenotypes" that go beyond single broad definition of disease and describe both healthy and sick patients. It is natural to

conjecture that the molecular basis of these "meta-phenotypes" will stem from new forms of GWAS, sequencing analysis, or other systems approaches using the molecular data collected through regular patient care.

Simple pairwise comparisons of diseases will no longer be sufficient for diagnosis and classification in the era of data-intensive medicine. A more scalable way to leverage these data is to conduct multiple studies on disease subclusters with specific comorbidity patterns to carefully tease out specific genetic and environmental determinants driving these disease subclusters[264, 130, 29]. Thus, it may be more plausible to conduct cheaper and targeted GWAS or sequencing analyses focusing on disease clusters with clear implication of genetic effects and little environmental noise, or exposure studies of strong environmental influences on disease subtypes regardless of genetic backgrounds [33, 262].

As we integrate many diverse data types into medical decision making, we become increasingly vulnerable to the incompleteness of our understanding on diseases. We could potentially underestimate the uncertainty when making medical decisions [190, 26]. The larger scope and size of the absolute information may inflate our confidence; compared to what we do know, what we do not know scales even faster.

# APPENDIX A

# SUPPLEMENTAL INFORMATION FOR "QUANTIFYING GENETIC AND ENVIRONMENTAL CONTRIBUTIONS ACROSS MULTIPLE DISEASES"

## A.1 Supplementary Note

### *A.1.1 Model Setup*

Assume that $m$ binary traits are observerd on $n$ individuals. The data for $i^{th}$ individual are $\mathbf{y_i}$, where $\mathbf{y_i} = (y_{i1}, \ldots, y_{im})$, and $i = 1, \ldots, n$. The observed value of each binary trait (such as disease or no disease) is assumed to be determined by an underlying Gaussian latent variable (called liability) represented by $\ell_{ij}$ for $j^{th}$ trait, $j \in \{1, \ldots, m\}$. For binary traits, we fixed the threshold at $t = 0$. It is assumed that $y_{ij} = 0$ iff $\ell_{ij} \leq 0$ and $y_{ij} = 1$ iff $\ell_{ij} > 0$. that is

$$y_{ij} = I(\ell_{ij} > 0) \tag{A.1}$$

where $I$ is the indicator function.

Define $\boldsymbol{\ell} = (\boldsymbol{\ell}_i)|_{i=1,\ldots,n}$ as the $nm$-dimensional vector containing the $\boldsymbol{\ell}_i$'s. Following the notations from [154], it is assumed that

$$\boldsymbol{\ell}|\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{R} \sim N_{nm}\Big(\mathbf{Xb} + \mathbf{Z_c c} + \mathbf{Z_a a}, \mathbf{R}\Big), \tag{A.2}$$

where $\mathbf{b}$ are the fixed effects , $\mathbf{a}$ are additive genetic effects, $\mathbf{c}$ are common environmental effects, and expected (co)variances of residual $\mathbf{R}$. If we assume phenotypes of different individuals are conditionally independent given parametes, but allow correlation between

residuals of phenotypes of the same individual. We can rewrite $\mathbf{R}$ as $\mathbf{I_n} \otimes \mathbf{R_e}$, where

$$\mathbf{R_e} = \begin{bmatrix} \sigma^2_{e,1} & \sigma_{e,12} \\ \sigma_{e,12} & \sigma^2_{e,2} \end{bmatrix} \tag{A.3}$$

## Incidence Matrices and Location Effects

Based on [265] we rewrite data $\mathbf{y}$ as $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$. For a bivariate model we can rewrite the model for $\boldsymbol{\ell}$ as:

$$\begin{bmatrix} \boldsymbol{\ell}_{y_1} \\ \boldsymbol{\ell}_{y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z_a}1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_a}2 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z_c}1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_c}2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}. \tag{A.4}$$

This is the same as (A.2), where fixed effects $\mathbf{b} = [\mathbf{b}'_1, \mathbf{b}'_2]'$, additive genetic effects $\mathbf{a} = [\mathbf{a}'_1, \mathbf{a}'_2]'$, common environmental effects $\mathbf{c} = [\mathbf{c}'_1, \mathbf{c}'_2]'$, and corresponding matrices $\mathbf{X}$, $\mathbf{Z_a}$, $\mathbf{Z_c}$, such that

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix}, \mathbf{Z_a} = \begin{bmatrix} \mathbf{Z_a}1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_a}2 \end{bmatrix} \mathbf{Z_c} = \begin{bmatrix} \mathbf{Z_c}1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_c}2 \end{bmatrix} \tag{A.5}$$

Furthermore we assumed $\mathbf{b_1}$ is a vector of order $p$, $\mathbf{a_1}$ is a vector of order $n$, and $\mathbf{c_1}$ is a vector of order $s$. Therefore $\mathbf{X}_1$ is a $n \times p$ matrix $\mathbf{Z_a}1$ is a $n \times n$ matrix and $\mathbf{Z_c}1$ is a $n \times s$ matrix. Given our model, we measured the same fixed effects for each trait( *e.g.* $\mathbf{X}_1 = \mathbf{X}_2$). Similarly, this is also true for $\mathbf{Z_a}$ and $\mathbf{Z_c}$ We further combined the location effects to vector $\boldsymbol{\theta} = [\mathbf{b}', \mathbf{c}', \mathbf{a}']'$ and incidence matrices to $\mathbf{W} = [\mathbf{X}, \mathbf{Z_a}, \mathbf{Z_c}]$

## Identifiability

In the multivariate Probit model, the parameters($\boldsymbol{\theta}$) are not identifiable from the observed-data [50]. This could be easily seen if we scale liability $\boldsymbol{\ell}$ by a constant $c > 0$

$$\begin{aligned} c\boldsymbol{\ell} &= c(\mathbf{w}'\boldsymbol{\theta} + \mathbf{e}) \\ &= \mathbf{w}'(c\boldsymbol{\theta}) + c\mathbf{e} \end{aligned} \tag{A.6}$$

Therefore, for binary data $\mathbf{y}$ will have the same value given $\boldsymbol{\theta}, \mathbf{e}$ and given $c\boldsymbol{\theta}, c\mathbf{e}$. [44] proposed that identifiability could be managed by restricting the covariance matrix to be a correlation matrix. Let $\mathbf{D} = \text{diag}(d_1, d_2), d_j > 0$ and define

$$\begin{aligned} \boldsymbol{\theta_j} &= d_j\tilde{\boldsymbol{\theta}_j} \\ \mathbf{R_e} &= \mathbf{D}\tilde{\mathbf{R}_e}\mathbf{D}' \end{aligned} \tag{A.7}$$

Follow the same rationale as (A.6), we have $\Pr(y_i|\mathbf{w}'\boldsymbol{\theta}, \mathbf{R_e}) = \Pr(y_i|\mathbf{w}'\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{R}_e})$. If we let

$$\mathbf{D} = \text{diag}(\tilde{\sigma}_{e,1}^{-1}, \tilde{\sigma}_{e,2}^{-1}) \tag{A.8}$$

$$\tilde{\mathbf{R}_e} = \begin{bmatrix} \tilde{\sigma}_{e,1}^2 & \tilde{\sigma}_{e,12} \\ \tilde{\sigma}_{e,12} & \tilde{\sigma}_{e,2}^2 \end{bmatrix} \tag{A.9}$$

$$\tag{A.10}$$

then we have

$$\mathbf{R_e} = \begin{bmatrix} \tilde{\sigma}_{e,1}^2\tilde{\sigma}_{e,1}^{-2} & \tilde{\sigma}_{e,12}/\tilde{\sigma}_{e,1}\tilde{\sigma}_{e,2} \\ \tilde{\sigma}_{e,12}/\tilde{\sigma}_{e,1}\tilde{\sigma}_{e,2} & \tilde{\sigma}_{e,2}^2\tilde{\sigma}_{e,2}^{-2} \end{bmatrix} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \tag{A.11}$$

where $r \in [-1, 1]$.

### *A.1.2 Prior distribution*

Starting with location effects, for the vector $\mathbf{b}$, follwing [108] we have

$$\mathbf{b}|\mathbf{B} \sim N_{mp}(\mathbf{0}, \mathbf{B}) \tag{A.12}$$

where $\mathbf{B} = \mathbf{I}_{mp}\sigma_b^2$ is the prior (co)variance matrix for the fixed effects with default value $\sigma_b^2 = 1\text{E}10$.

We assumed additive genetic effects follows,

$$\mathbf{a}|\mathbf{G} \sim N_{mn}(\mathbf{0}, \mathbf{G} \otimes \mathbf{A}) \tag{A.13}$$

where $\mathbf{A}$ is the additive genetic relationship matrix of order $n \times n$, and $\mathbf{G}$ is a $m \times m$ matrix , whose elements are the additive genetic (co)variance componenets. For bivariate model, we have

$$\mathbf{G} = \begin{bmatrix} \sigma_{a,1}^2 & \sigma_{a,12} \\ \sigma_{a,12} & \sigma_{a,2}^2 \end{bmatrix} \tag{A.14}$$

where the genetic correlation between traits, $\rho_{a,12} = \sigma_{a,12}/(\sigma_{a,1}\sigma_{a,2})$.

Similarly, We assumed common environment effects follows,

$$\mathbf{c}|\mathbf{R_c} \sim N_{ms}(\mathbf{0}, \mathbf{R_c} \otimes \mathbf{I_s}) \tag{A.15}$$

where $\mathbf{R_c}$ is a $m \times m$ matrix , whose elements are the common environment (co)variance componenets. For bivariate model, we have

$$\mathbf{R_c} = \begin{bmatrix} \sigma_{c,1}^2 & \sigma_{c,12} \\ \sigma_{c,12} & \sigma_{c,2}^2 \end{bmatrix} \tag{A.16}$$

The $m$-dimensional inverted Wishart distribtutions are assigned as prior for each of the $\mathbf{R_c}$ and $\mathbf{G}$ covariance matrices, with the respective densitites being

$$\mathbf{G} \sim IW_m(\nu_a, \mathbf{V}_a^{-1})) \tag{A.17}$$

$$\mathbf{R_c} \sim IW_m(\nu_c, \mathbf{V}_c^{-1})) \tag{A.18}$$

$$p(\mathbf{G}|\nu_a, V_a) \propto |\mathbf{G}|^{-\frac{1}{2}(\nu_a+m+1)} \exp[-\frac{1}{2} tr(\mathbf{G}^{-1}\mathbf{V}_a^{-1})] \tag{A.19}$$

$$p(\mathbf{R_c}|\nu_c, V_c) \propto |\mathbf{R_c}|^{-\frac{1}{2}(\nu_c+m+1)} \exp[-\frac{1}{2} tr(\mathbf{R_C}^{-1}\mathbf{V}_c^{-1})] \tag{A.20}$$

where $\mathbf{V}_a$ and $\nu_a$ are the prior sum of squares and prior degrees of freedom for additive genetic (co)variance matrix. And $\mathbf{V}_c$ and $\nu_c$ are the prior sum of squares and prior degree's of freedom for common environment (co)variance matrix

If we do not fixed $\mathbf{R}_e$ as an identity matrix and allow it to be a correlation matrix with the form

$$\mathbf{R_e} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \tag{A.21}$$

Following [174], $r$ following a prior distribution

$$p(r|\nu_e) \propto (1-r^2)^{-\frac{1}{2}(\nu_e+m+1)}, |r| < 1 \tag{A.22}$$

This is equivalent to the prior of $\mathbf{R}_e$ follows

$$p(\mathbf{R}_e|\nu_e) \propto |\mathbf{R}_e|^{-\frac{1}{2}(\nu_e+m+1)} I(\mathbf{R}_{e,jk} : \mathbf{R}_{e,jk} = 1(j=k), |\mathbf{R}_{e,jk}| < 1(j \neq k) \text{ and } \mathbf{R}_e \text{ is pos.def}) \tag{A.23}$$

### *A.1.3 Joint Posterior Density*

To simplify notations, we let

$$\boldsymbol{\Omega} = (\mathbf{b}, \mathbf{a}, \mathbf{c}, \mathbf{G}, \mathbf{R_c}, \mathbf{t}) \tag{A.24}$$

$$p(\boldsymbol{\Omega}) = p(\mathbf{b})p(\mathbf{a}|\mathbf{G})p(\mathbf{G})p(\mathbf{c}|\mathbf{R_c})p(\mathbf{R_c})p(\mathbf{t}) \tag{A.25}$$

Augmenting with the liability $\boldsymbol{\ell}$ we have

$$\begin{aligned} p(\boldsymbol{\Omega}, \boldsymbol{\ell}|\mathbf{y}) \propto & p(\mathbf{y}|\boldsymbol{\Omega}, \boldsymbol{\ell})p(\boldsymbol{\Omega}, \boldsymbol{\ell}) \\ = & p(\mathbf{y}|\boldsymbol{\ell}, \mathbf{t})p(\boldsymbol{\Omega}, \boldsymbol{\ell}) \\ = & p(\mathbf{y}|\boldsymbol{\ell}, \mathbf{t})p(\boldsymbol{\ell}|\boldsymbol{\theta})p(\boldsymbol{\Omega}) \end{aligned} \tag{A.26}$$

The probability that a given data point falls in a given category. given the liability and the thresholds ($p(\mathbf{y}|\boldsymbol{\ell}, \mathbf{t})$) is completed specified. For binary trait this is $[I(\ell_{ij} \leq 0)I(y_{ij} = 0) + I(\ell_{ij} > 0)I(y_{ij} = 1)]$. Combined with $p(\boldsymbol{\ell}|\boldsymbol{\theta})$ from (A.2), we have

$$\begin{aligned} p(\boldsymbol{\Omega}, \boldsymbol{\ell}|\mathbf{y}) \propto & \prod_{i=1}^{n} \Big[ \prod_{j=1}^{m} [I(\ell_{ij} \leq 0)I(y_{ij} = 0) + I(\ell_{ij} > 0)I(y_{ij} = 1)] \Big] \\ & \times \prod_{i=1}^{n} \left[ (2\pi)^{-\frac{m}{2}} |\mathbf{R}|^{-\frac{1}{2}} \exp(\{-\frac{1}{2}(\boldsymbol{\ell_i} - \mathbf{w}_i'\boldsymbol{\theta})'\mathbf{R}^{-1}(\boldsymbol{\ell_i} - \mathbf{w}_i'\boldsymbol{\theta})\}) \right] \\ & \times p(\mathbf{b})p(\mathbf{a}|\mathbf{G})p(\mathbf{G})p(\mathbf{c}|\mathbf{R_c})p(\mathbf{R_c})p(\mathbf{t}) \end{aligned} \tag{A.27}$$

### *A.1.4 Fully Conditional Posterior Distributions*

#### Liabilities

Given the model with threshold $\mathbf{t}$, we have

$$p(\boldsymbol{\ell}|\boldsymbol{\theta}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\ell}, \mathbf{t})p(\boldsymbol{\ell}|\boldsymbol{\theta}) \tag{A.28}$$

For binary trait with threshold fixed at 0 and conditionally on both the parameters and on data $\mathbf{y}_i$, $\boldsymbol{\ell}_i$ follows a truncated multivariate normal distribution:

$$\boldsymbol{\ell}_i|\boldsymbol{\theta}, \mathbf{y}_i \propto N_n(\mathbf{w}_i'\boldsymbol{\theta}, \mathbf{R}) \left[\prod_{j=1}^{m} [I(\ell_{ij} \leq 0)I(y_{ij} = 0) + I(\ell_{ij} > 0)I(y_{ij} = 1)]\right] \tag{A.29}$$

Since we assumed liabilities of the binary traits are conditionally independent, given parameters. Given the model with $\boldsymbol{\theta}$ and $\mathbf{t}$, we have

$$\Pr(y_{ij} = k|\boldsymbol{\theta}, \mathbf{t}) = \Pr(l_{ij} > t_k|\boldsymbol{\theta}) = \sum_{k=1}^{c} I(y_{ij} = k)[\Phi(t_k - \mathbf{w}_{ij}'\boldsymbol{\theta}) - \Phi(t_{k-1} - \mathbf{w}_{ij}'\boldsymbol{\theta})] \tag{A.30}$$

Where $\Phi$ and $\phi$ is the CDF and PDF of a standard normal variate. For binary triats, $t = 0$, this reduces to

$$\begin{aligned} \Pr(y_{ij} = 1|\boldsymbol{\theta}) =& \Phi(-(0 - \mathbf{w}_{ij}'\boldsymbol{\theta})) = \Phi(\mathbf{w}_{ij}'\boldsymbol{\theta}) \\ \Pr(y_{ij} = 0|\boldsymbol{\theta}) =& \Phi(-\mathbf{w}_{ij}'\boldsymbol{\theta}) \end{aligned} \tag{A.31}$$

Conditionally on both the parameters and on data $y_{ij}$, $l_{ij}$ follows a truncated normal distribution That is for $y_{ij} = 1$:

$$\Pr(\ell_{ij}|\boldsymbol{\theta}, y_{ij} = 1) = \frac{\phi(\mathbf{w}_{ij}'\boldsymbol{\theta}, 1)}{\Phi(\mathbf{w}_{ij}'\boldsymbol{\theta})} I(l_{ij} > 0) \tag{A.32}$$

Similarly for $y_{ij} = 0$, we have:

$$\Pr(\ell_{ij}|\boldsymbol{\theta}, y_{ij} = 0) = \frac{\phi(\mathbf{w}_{ij}'\boldsymbol{\theta}, 1)}{\Phi(-\mathbf{w}_{ij}'\boldsymbol{\theta})} I(l_{ij} \leq 0) \tag{A.33}$$

## Location Effects

The location effects($\boldsymbol{\theta} = [\mathbf{b}, \mathbf{a}, \mathbf{c}]$) and the residuals ($\mathbf{e}$) are assumed to come from a multivariate normal distribution:

$$\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \\ \mathbf{c} \\ \mathbf{e} \end{bmatrix} \sim N \left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{B} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \otimes \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R_c} \otimes \mathbf{I_s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R}_e \otimes \mathbf{I}_n \end{bmatrix} \right) \tag{A.34}$$

MCMCglmm [108] implemented a block sampler according to [88], $\boldsymbol{\theta}$ then can be sampled as a complete block by first solving the equations:

$$\tilde{\boldsymbol{\theta}} = \mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}(\boldsymbol{\ell} - \mathbf{W}\boldsymbol{\theta}_* - \mathbf{e}_*) \tag{A.35}$$

where $\mathbf{C}$ is the mixed model coefficient matrix:

$$\mathbf{C} = \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \begin{bmatrix} \mathbf{B}^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \otimes \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R_c}{-1} \otimes \mathbf{I_s} \end{bmatrix} \tag{A.36}$$

and $\boldsymbol{\theta}_*$ and $\mathbf{e}_*$ are random draws from the multivariate normal distributions:

$$\boldsymbol{\theta}_* \sim N \left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{B} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R_c} \otimes \mathbf{I_s} \end{bmatrix} \right) \tag{A.37}$$

$$\mathbf{e}_* \sim N\left(\mathbf{0}, \mathbf{R}_e \otimes \mathbf{I}_n\right) \tag{A.38}$$

Then $\tilde{\boldsymbol{\theta}} + \boldsymbol{\theta}_*$ gives a realisation from the probability distribution: $\Pr(\boldsymbol{\theta}|\boldsymbol{\ell}, \mathbf{W}, \mathbf{R_e}, \mathbf{R_c}, \mathbf{G}, \mathbf{A})$

## Dispersion Matrices

Here we continue with the fully conditional posterior distributions of the dispersion matrices. Starting with the genetic covariance matrix, from the joint posterior (A.27), we have:

$$p(\mathbf{G}|\mathbf{\Omega}, \mathbf{y}) \propto p(\mathbf{a}|\mathbf{G})p(\mathbf{G}) \tag{A.39}$$

From (A.13) we have

$$p(\mathbf{a}|\mathbf{G}) \propto |\mathbf{G}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i}^{n} \mathbf{a}_i'[\mathbf{G} \otimes \mathbf{A}]^{-1}\mathbf{a}_i\right] \tag{A.40}$$

where $\mathbf{a}_i = [a_{i1}, \cdots, a_{im}]$. Combining with (A.19) we have

$$\begin{aligned} p(\mathbf{G}|\mathbf{\Omega}, \mathbf{y}) \propto & p(\mathbf{a}|\mathbf{G})p(\mathbf{G}) \\ \propto & |\mathbf{G}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i}^{n} \mathbf{a}_i'[\mathbf{G} \otimes \mathbf{A}]^{-1}\mathbf{a}_i\right] \times |\mathbf{G}|^{-\frac{1}{2}(\nu_a+m+1)} \exp[-\frac{1}{2}tr(\mathbf{G}^{-1}\mathbf{V}_a^{-1})] \end{aligned} \tag{A.41}$$

Since a scalar is equal to the trace of itself, or $a = tr(a)$ (where $a$ is a scalar), which allows us to write $\sum_i^n \mathbf{a}_i'[\mathbf{G} \otimes \mathbf{A}]^{-1}\mathbf{a}_i$ as $tr(\sum_i^n \mathbf{a}_i'[\mathbf{G} \otimes \mathbf{A}]^{-1}\mathbf{a}_i)$ In addition, the trace operator is invariant under cyclic permutation, or $tr(AB) = tr(BA)$, which allows us to rotate our newly formed trace so that

$$\begin{aligned} tr\left[\sum_{i}^{n} \mathbf{a}_i'[\mathbf{G} \otimes \mathbf{A}]^{-1}\mathbf{a}_i\right] = & tr\left[\sum_{i}^{n} \mathbf{a}_i'\mathbf{a}_i[\mathbf{G} \otimes \mathbf{A}]^{-1}\right] = tr\left[(\sum_{i}^{n} \mathbf{a}_i'\mathbf{a}_i)[\mathbf{G} \otimes \mathbf{A}]^{-1}\right] \\ = & tr\left(\mathbf{S_a}[\mathbf{G} \otimes \mathbf{A}]^{-1}\right) = tr\left(\mathbf{G}^{-1}[\mathbf{S_a} \otimes \mathbf{A}^{-1}]\right) \end{aligned} \tag{A.42}$$

where $\mathbf{S_a}$ stands for the sums of squares matrix of $\mathbf{a}$. Together with the fact that the determinant of a product is the product of the determinants, or$|AB| = |A||B|$, we have

$$p(\mathbf{G}|\mathbf{\Omega}, \mathbf{y}) \propto |\mathbf{G}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_i^n \mathbf{a}_i'[\mathbf{G}\otimes\mathbf{A}]^{-1}\mathbf{a}_i\right] \times |\mathbf{G}|^{-\frac{1}{2}(\nu_a+m+1)} \exp[-\frac{1}{2}tr(\mathbf{G}^{-1}\mathbf{V}_a^{-1})] \tag{A.43}$$

$$\begin{aligned}
&= |\mathbf{G}|^{-\frac{n}{2}} |\mathbf{G}|^{-\frac{1}{2}(\nu_a+m+1)} \exp\left[-\frac{1}{2}tr\left(\mathbf{G}^{-1}[\mathbf{S_a}\otimes\mathbf{A}^{-1}]\right)\right] \exp\left[-\frac{1}{2}tr(\mathbf{G}^{-1}\mathbf{V}_a^{-1})\right] \\
&= |\mathbf{G}|^{-\frac{\nu_a+n+m+1}{2}} \exp\left[-\frac{1}{2}tr\left(\mathbf{G}^{-1}[\mathbf{S_a}\otimes\mathbf{A}^{-1}]\right) - \frac{1}{2}tr(\mathbf{G}^{-1}\mathbf{V}_a^{-1})\right] \\
&= |\mathbf{G}|^{-\frac{\nu_a+n+m+1}{2}} \exp\left[-\frac{1}{2}tr\left(\mathbf{G}^{-1}[\mathbf{S_a}\otimes\mathbf{A}^{-1}] + \mathbf{G}^{-1}\mathbf{V}_a^{-1}\right)\right] \\
&= |\mathbf{G}|^{-\frac{\nu_a+n+m+1}{2}} \exp\left[-\frac{1}{2}tr\left(\mathbf{G}^{-1}[\mathbf{S_a}\otimes\mathbf{A}^{-1} + \mathbf{V}_a^{-1}]\right)\right]
\end{aligned}$$

which indicates that $p(\mathbf{G}|\mathbf{\Omega}, \mathbf{y})$ follows a scaled inverse Wishart distribution:

$$p(\mathbf{G}|\mathbf{\Omega}, \mathbf{y}) = IW_m\left((\mathbf{S_a}\otimes\mathbf{A}^{-1} + \mathbf{V}_a^{-1})^{-1}, \nu_a + n\right) \tag{A.44}$$

Similarly, common environmental covariance matrix also follows a scaled inverse Wishart distribution:

$$p(\mathbf{R_c}|\mathbf{\Omega}, \mathbf{y}) = IW_m\left((\mathbf{S_c} + \mathbf{V}_c^{-1})^{-1}, \nu_c + s\right) \tag{A.45}$$

where $\mathbf{S_c} = \sum_i^s \mathbf{c}_i'\mathbf{c}_i$

For $\mathbf{R}_e$, as a correlation matrix, based on the algorithm provided by [15] we have

$$\begin{aligned}
p(\mathbf{R_e}|\mathbf{\Omega}, \mathbf{y}) &\propto |\mathbf{R}_e|^{-\frac{\nu_e+n+m+1}{2}} \exp\left[-\frac{1}{2}tr\left(\mathbf{R}_e^{-1}\mathbf{S}_e\right)\right] \\
p(\mathbf{R_e}|\mathbf{\Omega}, \mathbf{y}) &= IW_m\left(\mathbf{S}_e^{-1}, \nu_e + n\right)
\end{aligned} \tag{A.46}$$

where $\mathbf{S}_e = \sum_i^n \mathbf{e}_i'\mathbf{e}_i$ and $\mathbf{e}_i = \boldsymbol{\ell_i} - \mathbf{w}_i'\boldsymbol{\theta}$.

### *A.1.5 Variance Partition*

Here we focused on the estimations of three variance components, additive genetic (co)variance matrix $\mathbf{G}$, common environment (co)variance matrix $\mathbf{R_c}$ and unique environment (co)variance matrix $\mathbf{R_e}$. We have

$$\mathbf{G} = \begin{bmatrix} \sigma^2_{a,1} & \sigma_{a,12} \\ \sigma_{a,12} & \sigma^2_{a,2} \end{bmatrix}, \mathbf{R_c} = \begin{bmatrix} \sigma^2_{c,1} & \sigma_{c,12} \\ \sigma_{c,12} & \sigma^2_{c,2} \end{bmatrix}, \text{and } \mathbf{R_e} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \tag{A.47}$$

where $\sigma_{a,12}$ is the additive genetic variance that influences both traits, $\sigma_{c,12}$ is the common environment variance that influences both traits, and $r$ is the unique environment variance that influences both traits. The narrow-sense heritability of trait $j$ on the liability scaleof trait $j$ is

$$h^2_j = \frac{\sigma^2_{a,j}}{\sigma^2_{a,j} + \sigma^2_{c,j} + 1} \tag{A.48}$$

## Shared variance

The genetic correlation between traits is defined as $\rho_{a,12} = \sigma_{a,12}/(\sigma_{a,1}\sigma_{a,2})$. And $\rho^2_{a,12}$ is the proportion of covariability in the total variability of two traits, where

$$\rho^2_{a,12} = (\frac{\sigma_{a,12}}{\sigma_{a,1}\sigma_{a,2}})^2 = \frac{|\sigma_{a,12}|}{\sigma^2_{a,1}} \frac{|\sigma_{a,12}|}{\sigma^2_{a,2}} \tag{A.49}$$

If we consider $prop_1 = \frac{|\sigma_{a,12}|}{\sigma^2_{a,1}}$ as the proportion of the shared genetic variance for trait 1 and $prop_2 = \frac{|\sigma_{a,12}|}{\sigma^2_{a,2}}$ as the proportion of the shared genetic variance for trait 2, the genetic correlation $\rho_{a,12}$ could be thought of as their geometric mean, $\sqrt{prop_1 \times prop_2}$. Similarly we can calculate proportion of shared variance for common environment variance and unique environment variance.

## Co-heritability

According to [74], the liability is measured on the scale of the standard deviation of its distribution $\sqrt{V_{P,j}}$, therefore we have heritability as follows:

$$h_j^2 = \frac{V_{a,j}}{V_{P,j}} = \frac{V_{a,j}}{V_{a,j} + V_{c,j} + V_{e,j}} \tag{A.50}$$

Because of the identification, we choose to fix $\sigma_{e,j}^2 = 1$. Therefore we have

$$h_j^2 = \frac{V_{a,j}/V_{e,j}}{V_{P,j}/V_{e,j}} = \frac{\sigma_{a,j}^2}{\sigma_{P,j}^2} = \frac{\sigma_{a,j}^2}{\sigma_{a,j}^2 + \sigma_{c,j}^2 + 1} \tag{A.51}$$

Extending the notion of heritability on the liability scale, coheritability of two diseases $i$ and $j$ should be measured on the scale of the product of the two standard deviations $\sqrt{V_{P,i}}\sqrt{V_{P,j}}$

$$coh_{ij} = \frac{Cov_{a,ij}}{\sqrt{V_{P,i}}\sqrt{V_{P,j}}} = \frac{Cov_{a,ij}/\sqrt{V_{e,i}}\sqrt{V_{e,j}}}{\sqrt{V_{P,i}/V_{e,i}}\sqrt{V_{P,j}/V_{e,j}}} = \frac{\sigma_{a,ij}}{\sigma_{P,i}\sigma_{P,j}} = \rho_{a,ij}h_i h_j \tag{A.52}$$

$$h_j^2 = \frac{V_{a,j}}{V_{P,j}} = \frac{\mathrm{cov}(a_j, p_j)}{\sqrt{V_{P,j}}\sqrt{V_{P,j}}} = \frac{\mathrm{cov}(a_j, a_j + c_j + e_j)}{\sqrt{V_{P,j}}\sqrt{V_{P,j}}} = \frac{\mathrm{cov}(a_j, a_j)}{\sqrt{V_{P,j}}\sqrt{V_{P,j}}} = \frac{Var(a_j, a_j)}{\sqrt{V_{P,j}}\sqrt{V_{P,j}}} \tag{A.53}$$

$$coh_{ij} = \frac{\mathrm{cov}(a_i, p_j)}{\sqrt{V_{P,i}}\sqrt{V_{P,j}}} = \frac{\mathrm{cov}(a_i, a_j + c_j + e_j)}{\sqrt{V_{P,i}}\sqrt{V_{P,j}}} = \frac{\mathrm{cov}(a_i, a_j)}{\sqrt{V_{P,i}}\sqrt{V_{P,j}}} \tag{A.54}$$

$coh_{ij}$ represents how much effect does the genetic factor of one trait has on the other trait.

$$\text{co-h}_{ij} = \frac{\mathrm{cov}(a_i, p_j)}{\mathrm{cov}(p_i, p_j)} = \frac{\mathrm{cov}(a_i, a_j)}{\mathrm{cov}(a_i, a_j) + \mathrm{cov}(e_i, e_j)} \tag{A.55}$$

co-h$_{ij}$ represents how much effect does genetic factor on the covariance of two traits. Heritability reflects the degree to which the genes transmitted from the parents determine the trait of their children. $coh_{ij}$ represents the degree to which the genes transmitted from the parents that determine one trait that can also determine the second trait. And co-h$_{ij}$ represents the degree to which the genes transmitted from the parents determine the covariance of two traits.

## A.2 Supplementary Figures

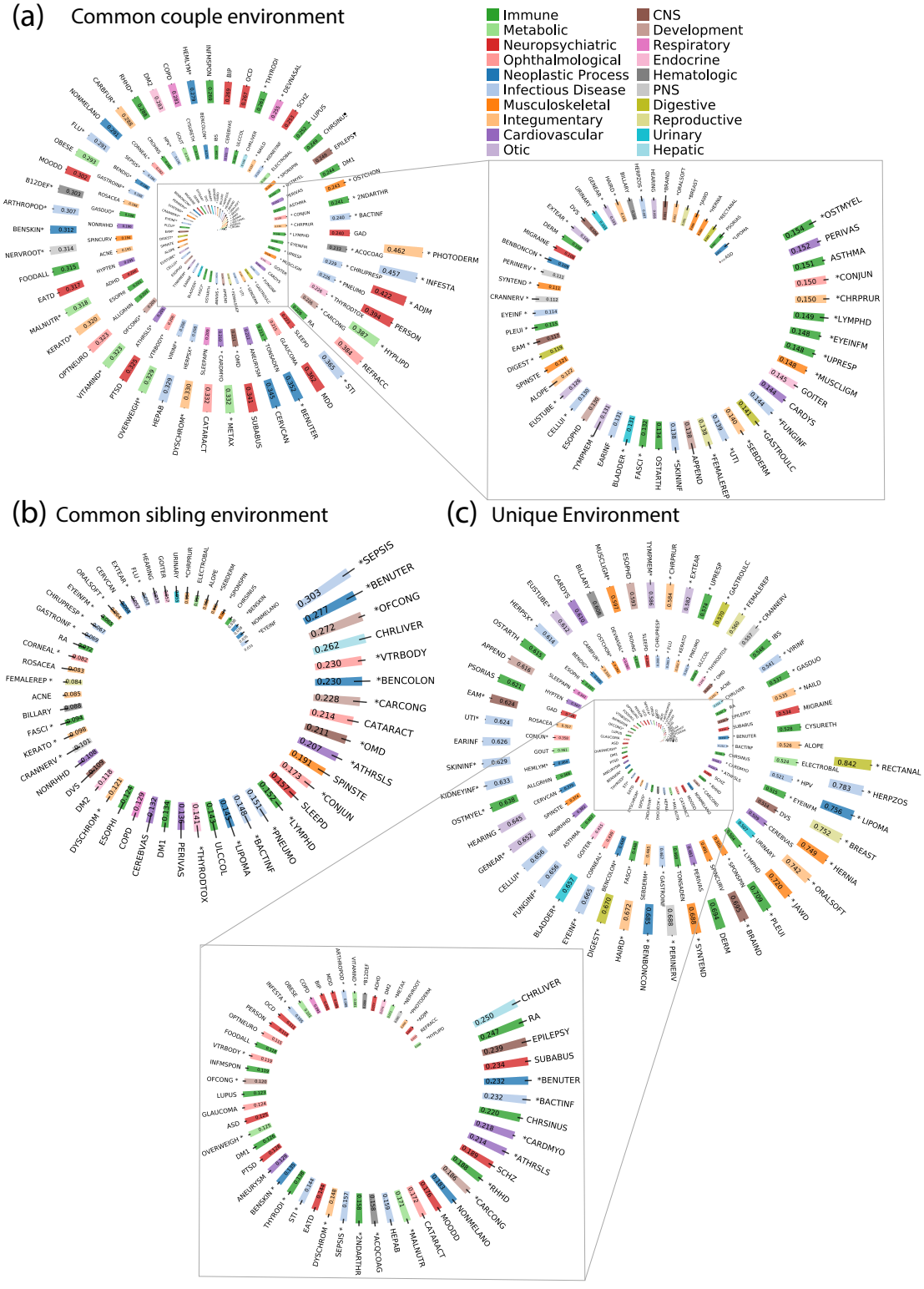


**Figure A.1: Testing dependence of heritability estimates on age of onset; heritability distributions, sorted by biological system.**

**Figure A.1 Continued** (A) Histograms and density plots of heritability estimates by biological system. (B) Heritability estimate vs. disease age of onset for biological system with more than 3 diseases, linear fits indicated by solid lines.

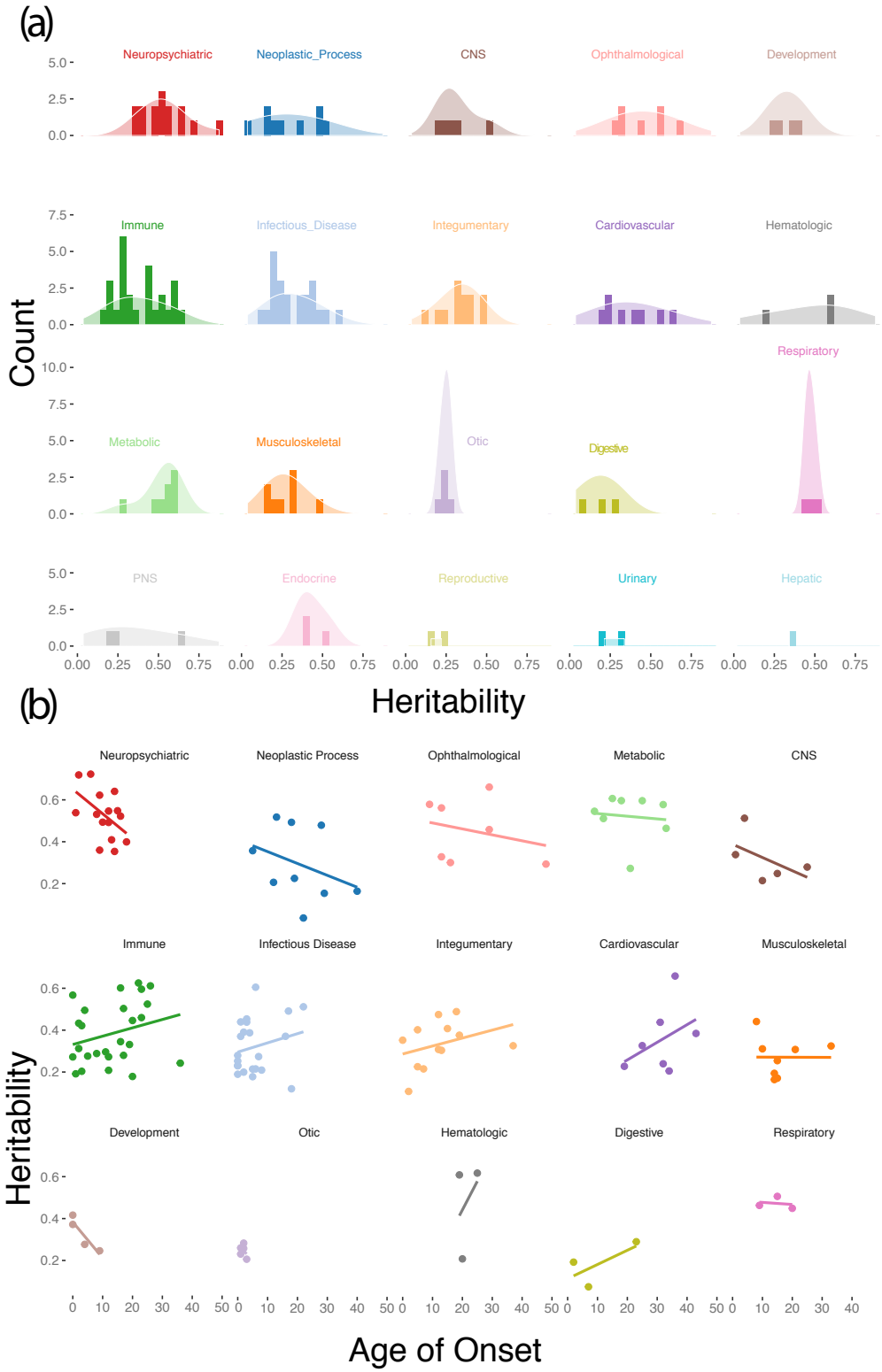


**Figure A.2: Environmental effects estimates.**

**Figure A.2 Continued** (A) Common couple environment effects. (B) Common Sibling environment effects. (C) Unique environment effects. Bar color in the bar plots indicates biological systems associated with each disease, consistent throughout all figures.

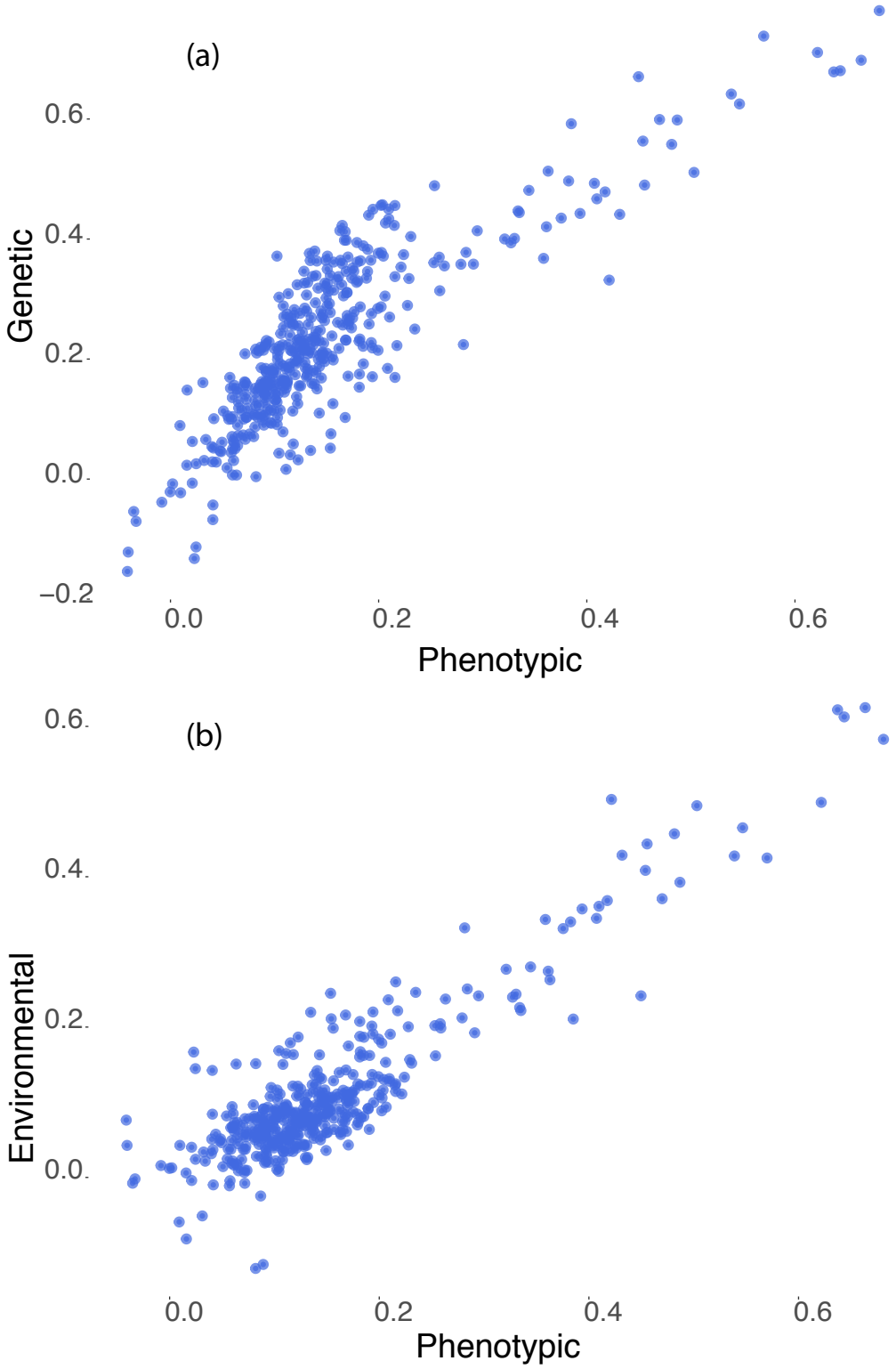


**Figure A.3: Positive correlations between phenotypic and genetic correlations** (A) and phenotypic and environmental correlations (B).

**Figure A.3 Continued Positive correlations between phenotypic and genetic correlations** (A) and phenotypic and environmental correlations (B).

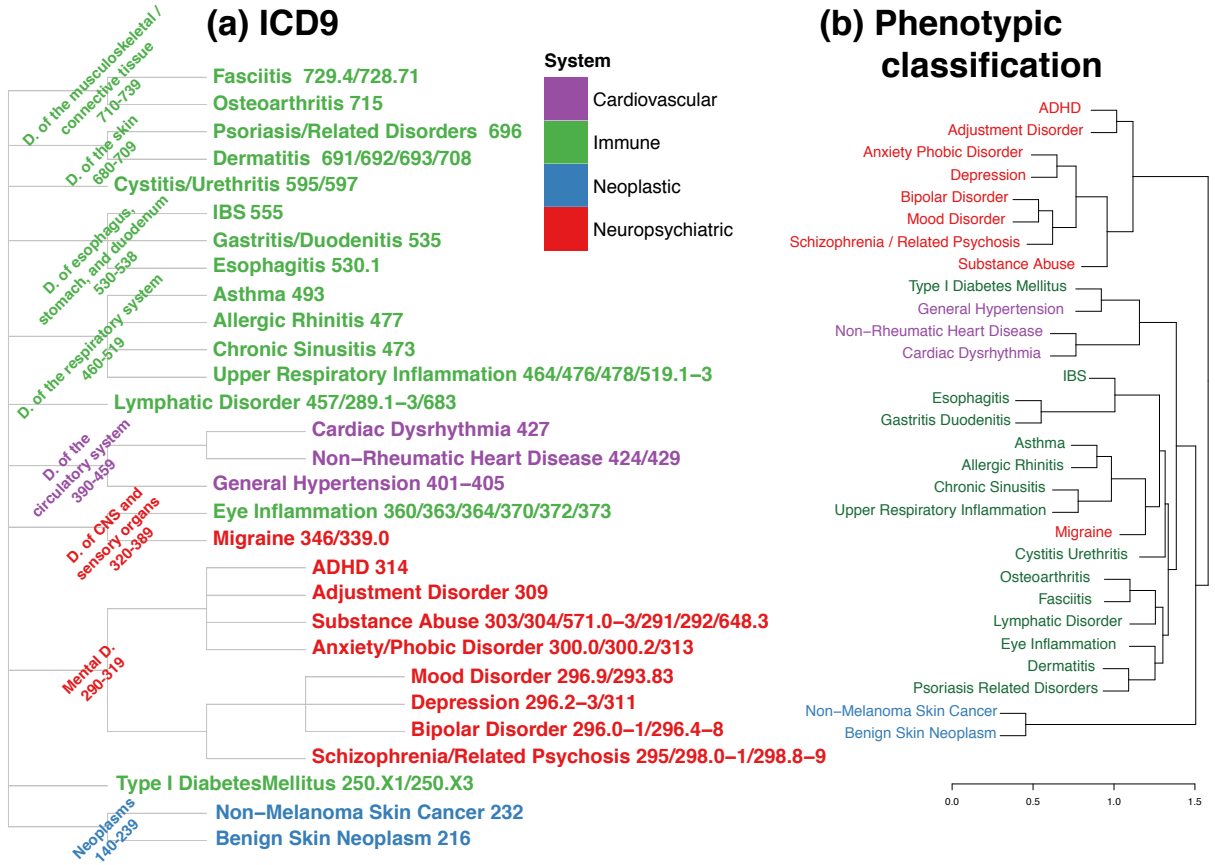


**Figure A.4: Classification trees: ICD9 vs phenotypic correlations.** (A) A classification of diseases that corresponds to a subset of ICD9 taxonomy. (B) Disease classification constructed from phenotypic correlations between diseases; distances between diseases were calculated as 1—correlation.

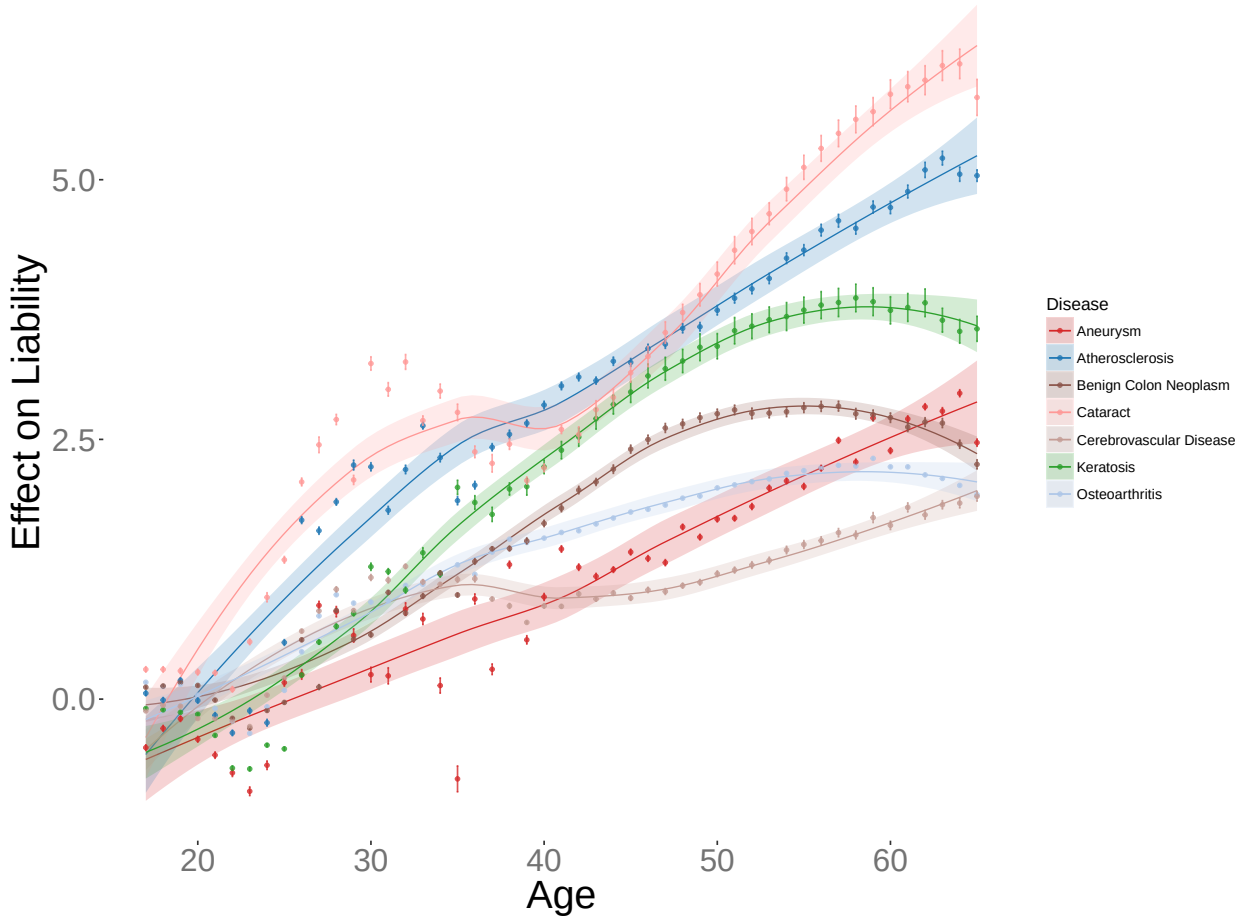


**Figure A.5: Estimates of age-related increase in disease liability for seven late-onset conditions (aneurysm, atherosclerosis , benign colon neoplasm, cataract, cerebrovascular disease, keratosis , and osteoarthritis).** Error bars show one standard deviation, and LOcally WEighted Scatter-plot Smoother (LOWESS) curve fits are shown with solid lines.

## A.3 Supplementary Tables

**Table A.1:** Acronyms, biological systems, prevalence percentages and standard errors for 149 studied diseases

**Supplementary Table 1. Acronyms, biological systems, prevalence percentages and standard errors for 149 studied diseases**

| Disease | Acronym | System | Age of Onset | All Patient | Prevalence | Prevalence SE | Total Study | Prevalence in | SE in Study | Parents in | Prevalence in | SE in Parent | Children in | Prevalence in | SE in Children |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Brain Damage | BRAIND | CNS | 10 | 781369 | 0.52% | 4.2E-06 | 5737 | 1.19% | 0.016% | 3314 | 1.28% | 0.016% | 2423 | 1.08% | 0.015% |
| Disorder of the Vestibular System | DVS | CNS | 25 | 1432063 | 0.95% | 5.7E-06 | 12356 | 2.57% | 0.023% | 10782 | 4.18% | 0.029% | 1574 | 0.70% | 0.012% |
| Epilepsy Related Disorders | EPILEPSY | CNS | 4 | 937933 | 0.62% | 4.6E-06 | 8345 | 1.73% | 0.019% | 3097 | 1.20% | 0.016% | 5248 | 2.35% | 0.022% |
| Extrapyramidal Abnormal Movement Disorders | EAM | CNS | 15 | 1201840 | 0.80% | 5.2E-06 | 9372 | 1.95% | 0.020% | 7151 | 2.77% | 0.024% | 2221 | 0.99% | 0.014% |
| Ocular Musculoskeletal Disorder | OMD | CNS | 1 | 1646327 | 1.09% | 6.1E-06 | 10069 | 2.09% | 0.021% | 6369 | 2.47% | 0.022% | 3700 | 1.65% | 0.018% |
| Aneurysm | ANEURYSM | Cardiovascular | 36 | 933165 | 0.62% | 4.6E-06 | 5347 | 1.11% | 0.015% | 4895 | 1.90% | 0.020% | 452 | 0.20% | 0.006% |
| Atherosclerosis | ATHRSLS | Cardiovascular | 43 | 6134419 | 4.06% | 1.2E-05 | 32595 | 6.77% | 0.036% | 31564 | 12.24% | 0.047% | 1031 | 0.46% | 0.010% |
| Cardiac Dysrhythmia | CARDYS | Cardiovascular | 19 | 6749062 | 4.47% | 1.2E-05 | 43391 | 9.01% | 0.041% | 35054 | 13.59% | 0.049% | 8337 | 3.73% | 0.027% |
| Cardiomyopathy | CARDMYO | Cardiovascular | 32 | 1015249 | 0.67% | 4.8E-06 | 5436 | 1.13% | 0.015% | 4788 | 1.86% | 0.019% | 648 | 0.29% | 0.008% |
| Cerebrovascular Disease | CEREBVAS | Cardiovascular | 34 | 4201911 | 2.78% | 9.6E-06 | 22844 | 4.74% | 0.031% | 20673 | 8.01% | 0.039% | 2171 | 0.97% | 0.014% |
| General Hypertension | HYPTEN | Cardiovascular | 31 | 26193165 | 17.33% | 2.2E-05 | 142283 | 29.54% | 0.066% | 1E+05 | 50.71% | 0.072% | 11471 | 5.13% | 0.032% |
| Non-Rheumatic Heart Disease | NONRHHD | Cardiovascular | 25 | 4513115 | 2.99% | 9.9E-06 | 29734 | 6.17% | 0.035% | 25291 | 9.80% | 0.043% | 4443 | 1.99% | 0.020% |
| Peripheral Vascular Disease | PERIVAS | Cardiovascular | 32 | 2846560 | 1.88% | 7.9E-06 | 17417 | 3.62% | 0.027% | 15530 | 6.02% | 0.034% | 1887 | 0.84% | 0.013% |
| Cardiac Congenital Anomaly | CARCONG | Development | 0 | 862287 | 0.57% | 4.4E-06 | 4844 | 1.01% | 0.014% | 3016 | 1.17% | 0.015% | 1828 | 0.82% | 0.013% |
| Oro-Facial Congenital Anomaly | OFCONG | Development | 0 | 1332387 | 0.88% | 5.5E-06 | 15151 | 3.15% | 0.025% | 1265 | 0.49% | 0.010% | 13886 | 6.21% | 0.035% |
| Appendiceal Disease | APPEND | Development | 9 | 560036 | 0.37% | 3.5E-06 | 5243 | 1.09% | 0.015% | 2584 | 1.00% | 0.014% | 2659 | 1.19% | 0.016% |
| Esophageal Disease | ESOPHD | Development | 4 | 11623811 | 7.69% | 1.6E-05 | 80534 | 16.72% | 0.054% | 63315 | 24.54% | 0.062% | 17219 | 7.70% | 0.038% |
| Functional Digestive Disorder | DIGEST | Digestive | 2 | 7738872 | 5.12% | 1.3E-05 | 52203 | 10.84% | 0.045% | 37011 | 14.35% | 0.051% | 15192 | 6.79% | 0.036% |
| Gastrointestinal Ulcer | GASTROULC | Digestive | 23 | 1190903 | 0.79% | 5.2E-06 | 9305 | 1.93% | 0.020% | 7510 | 2.91% | 0.024% | 1795 | 0.80% | 0.013% |
| Rectal Anal Disorder | RECTANAL | Digestive | 7 | 1401511 | 0.93% | 5.6E-06 | 12456 | 2.59% | 0.023% | 8925 | 3.46% | 0.026% | 3531 | 1.58% | 0.018% |
| Goiter | GOITER | Endocrine | 21 | 2453163 | 1.62% | 7.4E-06 | 20030 | 4.16% | 0.029% | 15759 | 6.11% | 0.035% | 4271 | 1.91% | 0.020% |
| Thyrotoxicosis | THYRODTOX | Endocrine | 21 | 1005843 | 0.67% | 4.7E-06 | 8910 | 1.85% | 0.019% | 6705 | 2.60% | 0.023% | 2205 | 0.99% | 0.014% |
| Type II Diabetes Mellitus | DM2 | Endocrine | 31 | 9943788 | 6.58% | 1.4E-05 | 53720 | 11.15% | 0.045% | 48206 | 18.69% | 0.056% | 5514 | 2.47% | 0.022% |
| Acquired Coagulation Defect | ACQCOAG | Hematologic | 19 | 550859 | 0.36% | 3.5E-06 | 4268 | 0.89% | 0.014% | 3324 | 1.29% | 0.016% | 944 | 0.42% | 0.009% |
| Pernicious or B12 Deficiency Anemia | B12DEF | Hematologic | 25 | 1243817 | 0.82% | 5.3E-06 | 7886 | 1.64% | 0.018% | 6479 | 2.51% | 0.023% | 1407 | 0.63% | 0.011% |
| Biliary Tract Disease | BILLARY | Hematologic | 20 | 3472467 | 2.30% | 8.7E-06 | 29647 | 6.16% | 0.035% | 22120 | 8.57% | 0.040% | 7527 | 3.37% | 0.026% |
| Chronic Liver Disease | CHRLIVER | Hepatic | 27 | 1654780 | 1.10% | 6.1E-06 | 13154 | 2.73% | 0.023% | 11544 | 4.47% | 0.030% | 1610 | 0.72% | 0.012% |
| Allergic Rhinitis | ALLGRHIN | Immune | 3 | 16282209 | 10.78% | 1.8E-05 | 106606 | 22.13% | 0.060% | 61707 | 23.92% | 0.061% | 44899 | 20.07% | 0.058% |
| Asthma | ASTHMA | Immune | 2 | 9455688 | 6.26% | 1.4E-05 | 52764 | 10.95% | 0.045% | 28680 | 11.12% | 0.045% | 24084 | 10.77% | 0.045% |
| Atopic Contact Dermatitis | DERM | Immune | 1 | 14336386 | 9.49% | 1.7E-05 | 103205 | 21.43% | 0.059% | 62729 | 24.32% | 0.062% | 40476 | 18.10% | 0.055% |
| Chronic Sinusitis | CHRSINUS | Immune | 4 | 7150277 | 4.73% | 1.2E-05 | 59361 | 12.32% | 0.047% | 38603 | 14.96% | 0.051% | 20758 | 9.28% | 0.042% |
| Crohns Disease | CROHNS | Immune | 15 | 435641 | 0.29% | 3.1E-06 | 3485 | 0.72% | 0.012% | 2151 | 0.83% | 0.013% | 1334 | 0.60% | 0.011% |
| Chronic Tonsillitis Adenoiditis | TONSADEN | Immune | 2 | 1742103 | 1.15% | 6.2E-06 | 10865 | 2.26% | 0.021% | 1990 | 0.77% | 0.013% | 8875 | 3.97% | 0.028% |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cystitis Urethritis | CYSURETH | Immune | 11 | 2132075 | 1.41% | 6.9E-06 | 19675 | 4.08% | 0.029% | 11259 | 4.36% | 0.029% | 8416 | 3.76% | 0.027% |
| Esophagitis | ESOPHI | Immune | 16 | 1558769 | 1.03% | 5.9E-06 | 14594 | 3.03% | 0.025% | 11617 | 4.50% | 0.030% | 2977 | 1.33% | 0.017% |
| Eye Inflammation | EYEINFM | Immune | 5 | 6867051 | 4.54% | 1.2E-05 | 57174 | 11.87% | 0.047% | 36358 | 14.09% | 0.050% | 20816 | 9.31% | 0.042% |
| Fasciitis | FASCI | Immune | 19 | 3340464 | 2.21% | 8.6E-06 | 33397 | 6.93% | 0.037% | 29209 | 11.32% | 0.046% | 4188 | 1.87% | 0.020% |
| Food Allergy | FOODALL | Immune | 0 | 661417 | 0.44% | 3.9E-06 | 3730 | 0.77% | 0.013% | 2042 | 0.79% | 0.013% | 1688 | 0.75% | 0.012% |
| Gastritis Duodenitis | GASDUO | Immune | 12 | 5083925 | 3.36% | 1.1E-05 | 42044 | 8.73% | 0.041% | 30230 | 11.72% | 0.046% | 11814 | 5.28% | 0.032% |
| IBS | IBS | Immune | 16 | 1903806 | 1.26% | 6.5E-06 | 18069 | 3.75% | 0.027% | 11409 | 4.42% | 0.030% | 6660 | 2.98% | 0.024% |
| Inflammatory Spondylopathies | INFMSPON | Immune | 26 | 716012 | 0.47% | 4.0E-06 | 6111 | 1.27% | 0.016% | 5333 | 2.07% | 0.021% | 778 | 0.35% | 0.008% |
| Lupus Erythematosus | LUPUS | Immune | 22 | 378480 | 0.25% | 2.9E-06 | 2883 | 0.60% | 0.011% | 2234 | 0.87% | 0.013% | 649 | 0.29% | 0.008% |
| Lymphatic Disorder | LYMPHD | Immune | 3 | 1060131 | 0.70% | 4.9E-06 | 8347 | 1.73% | 0.019% | 4599 | 1.78% | 0.019% | 3748 | 1.68% | 0.018% |
| Osteoarthritis | OSTARTH | Immune | 36 | 10204522 | 6.75% | 1.5E-05 | 76703 | 15.92% | 0.053% | 72296 | 28.02% | 0.065% | 4407 | 1.97% | 0.020% |
| Osteomyelitis | OSTMYEL | Immune | 12 | 1059798 | 0.70% | 4.9E-06 | 8555 | 1.78% | 0.019% | 5135 | 1.99% | 0.020% | 3420 | 1.53% | 0.018% |
| Pleuritis | PLEUI | Immune | 20 | 1855863 | 1.23% | 6.4E-06 | 10835 | 2.25% | 0.021% | 8220 | 3.19% | 0.025% | 2615 | 1.17% | 0.015% |
| Psoriasis Related Disorders | PSORIAS | Immune | 8 | 1625758 | 1.08% | 6.0E-06 | 12858 | 2.67% | 0.023% | 7955 | 3.08% | 0.025% | 4903 | 2.19% | 0.021% |
| Rheumatic Heart Disease | RHHD | Immune | 25 | 1130316 | 0.75% | 5.0E-06 | 6845 | 1.42% | 0.017% | 5794 | 2.25% | 0.021% | 1051 | 0.47% | 0.010% |
| Rheumatoid Arthritis Related Conditions | RA | Immune | 23 | 1380829 | 0.91% | 5.6E-06 | 9897 | 2.05% | 0.020% | 8452 | 3.28% | 0.026% | 1445 | 0.65% | 0.012% |
| Secondary Arthropathy | 2NDARTHR | Immune | 23 | 794542 | 0.53% | 4.2E-06 | 6316 | 1.31% | 0.016% | 5239 | 2.03% | 0.020% | 1077 | 0.48% | 0.010% |
| Thyroiditis | THYRODI | Immune | 16 | 845338 | 0.56% | 4.4E-06 | 7595 | 1.58% | 0.018% | 5616 | 2.18% | 0.021% | 1979 | 0.88% | 0.013% |
| Type I Diabetes Mellitus | DM1 | Immune | 17 | 1822886 | 1.21% | 6.4E-06 | 10711 | 2.22% | 0.021% | 8689 | 3.37% | 0.026% | 2022 | 0.90% | 0.014% |
| Ulcerative Colitis | ULCCOL | Immune | 20 | 522581 | 0.35% | 3.4E-06 | 4148 | 0.86% | 0.013% | 2964 | 1.15% | 0.015% | 1184 | 0.53% | 0.010% |
| Upper Respiratory Inflammation | UPRESP | Immune | 0 | 4511399 | 2.99% | 9.9E-06 | 33089 | 6.87% | 0.036% | 21492 | 8.33% | 0.040% | 11597 | 5.18% | 0.032% |
| Arthropod-Borne Diseases | ARTHROPOD | Infectious Diseas | 6 | 310003 | 0.21% | 2.6E-06 | 2423 | 0.50% | 0.010% | 1666 | 0.65% | 0.012% | 757 | 0.34% | 0.008% |
| Cellulitis | CELLUI | Infectious Diseas | 5 | 9165144 | 6.07% | 1.4E-05 | 71010 | 14.74% | 0.051% | 43709 | 16.94% | 0.054% | 27301 | 12.21% | 0.047% |
| Chronic Upper Respiratory Infection | CHRUPRESP | Infectious Diseas | 1 | 2883109 | 1.91% | 8.0E-06 | 20521 | 4.26% | 0.029% | 13280 | 5.15% | 0.032% | 7241 | 3.24% | 0.026% |
| Ear Infection | EARINF | Infectious Diseas | 0 | 16090384 | 10.65% | 1.8E-05 | 72674 | 15.09% | 0.052% | 39334 | 15.25% | 0.052% | 33340 | 14.91% | 0.051% |
| Eye Infection | EYEINF | Infectious Diseas | 0 | 8006448 | 5.30% | 1.3E-05 | 53095 | 11.02% | 0.045% | 29276 | 11.35% | 0.046% | 23819 | 10.65% | 0.044% |
| Fungal Infection | FUNGINF | Infectious Diseas | 2 | 9524283 | 6.30% | 1.4E-05 | 68836 | 14.29% | 0.050% | 41720 | 16.17% | 0.053% | 27116 | 12.12% | 0.047% |
| Gastrointestinal Infection | GASTROINF | Infectious Diseas | 0 | 2810029 | 1.86% | 7.9E-06 | 16104 | 3.34% | 0.026% | 8426 | 3.27% | 0.026% | 7678 | 3.43% | 0.026% |
| General Bacterial Infection | BACTINF | Infectious Diseas | 4 | 704538 | 0.47% | 4.0E-06 | 4645 | 0.96% | 0.014% | 2663 | 1.03% | 0.015% | 1982 | 0.89% | 0.014% |
| General Viral Infection | VIRINF | Infectious Diseas | 0 | 8453411 | 5.59% | 1.3E-05 | 37702 | 7.83% | 0.039% | 14966 | 5.80% | 0.034% | 22736 | 10.16% | 0.044% |
| Herpes Simplex | HERPSX | Infectious Diseas | 5 | 1734639 | 1.15% | 6.2E-06 | 13706 | 2.85% | 0.024% | 6939 | 2.69% | 0.023% | 6767 | 3.03% | 0.025% |
| Herpes Zoster | HERPZOS | Infectious Diseas | 18 | 1506363 | 1.00% | 5.8E-06 | 13362 | 2.77% | 0.024% | 10659 | 4.13% | 0.029% | 2703 | 1.21% | 0.016% |
| Infestation | INFESTA | Infectious Diseas | 3 | 549056 | 0.36% | 3.5E-06 | 3905 | 0.81% | 0.013% | 1498 | 0.58% | 0.011% | 2407 | 1.08% | 0.015% |
| Influenza | FLU | Infectious Diseas | 2 | 3917728 | 2.59% | 9.3E-06 | 23166 | 4.81% | 0.031% | 11764 | 4.56% | 0.030% | 11402 | 5.10% | 0.032% |
| Kidney Infection | KIDNEYINF | Infectious Diseas | 8 | 752345 | 0.50% | 4.1E-06 | 7098 | 1.47% | 0.017% | 3354 | 1.30% | 0.016% | 3744 | 1.67% | 0.018% |
| Pneumonia | PNEUMO | Infectious Diseas | 1 | 1102652 | 0.73% | 5.0E-06 | 5654 | 1.17% | 0.016% | 3950 | 1.53% | 0.018% | 1704 | 0.76% | 0.013% |
| STI | STI | Infectious Diseas | 17 | 500652 | 0.33% | 3.4E-06 | 5100 | 1.06% | 0.015% | 1100 | 0.43% | 0.009% | 4000 | 1.79% | 0.019% |
| Septicemia | SEPSIS | Infectious Diseas | 16 | 862110 | 0.57% | 4.4E-06 | 4486 | 0.93% | 0.014% | 3242 | 1.26% | 0.016% | 1244 | 0.56% | 0.011% |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Skin Infection | SKININF | Infectious Diseas | 0 | 1776635 | 1.18% | 6.3E-06 | 15350 | 3.19% | 0.025% | 8494 | 3.29% | 0.026% | 6856 | 3.07% | 0.025% |
| UTI | UTI | Infectious Diseas | 6 | 12538001 | 8.30% | 1.6E-05 | 98197 | 20.39% | 0.058% | 59412 | 23.03% | 0.061% | 38785 | 17.34% | 0.055% |
| Viral Hepatitis B | HEPAB | Infectious Diseas | 22 | 155890 | 0.10% | 1.9E-06 | 977 | 0.20% | 0.006% | 765 | 0.30% | 0.008% | 212 | 0.09% | 0.004% |
| Viral Warts HPV | HPV | Infectious Diseas | 7 | 5695793 | 3.77% | 1.1E-05 | 55223 | 11.47% | 0.046% | 24426 | 9.47% | 0.042% | 30797 | 13.77% | 0.050% |
| Acne | ACNE | Integumentary | 12 | 5490739 | 3.63% | 1.1E-05 | 60005 | 12.46% | 0.048% | 9638 | 3.74% | 0.027% | 50367 | 22.52% | 0.060% |
| Alopecia | ALOPE | Integumentary | 13 | 1066362 | 0.71% | 4.9E-06 | 9330 | 1.94% | 0.020% | 5366 | 2.08% | 0.021% | 3964 | 1.77% | 0.019% |
| Carbuncle Furuncle | CARBFUR | Integumentary | 5 | 708165 | 0.47% | 4.0E-06 | 6299 | 1.31% | 0.016% | 3308 | 1.28% | 0.016% | 2991 | 1.34% | 0.017% |
| Chronic Pruritus | CHRPRUR | Integumentary | 7 | 1568834 | 1.04% | 5.9E-06 | 11797 | 2.45% | 0.022% | 8099 | 3.14% | 0.025% | 3698 | 1.65% | 0.018% |
| Dyschromia | DYSCHROM | Integumentary | 15 | 2888863 | 1.91% | 8.0E-06 | 24653 | 5.12% | 0.032% | 20300 | 7.87% | 0.039% | 4353 | 1.95% | 0.020% |
| Hair Related Disease | HAIRD | Integumentary | 5 | 1857982 | 1.23% | 6.4E-06 | 18554 | 3.85% | 0.028% | 8585 | 3.33% | 0.026% | 9969 | 4.46% | 0.030% |
| Keratosis | KERATO | Integumentary | 37 | 8695184 | 5.75% | 1.4E-05 | 77326 | 16.05% | 0.053% | 75200 | 29.15% | 0.065% | 2126 | 0.95% | 0.014% |
| Nail Disease | NAILD | Integumentary | 12 | 2888074 | 1.91% | 8.0E-06 | 22506 | 4.67% | 0.030% | 14584 | 5.65% | 0.033% | 7922 | 3.54% | 0.027% |
| Oral Soft Tissue Disease | ORALSOFT | Integumentary | 2 | 1602926 | 1.06% | 6.0E-06 | 13679 | 2.84% | 0.024% | 8241 | 3.19% | 0.025% | 5438 | 2.43% | 0.022% |
| Photodermatitis | PHOTODERM | Integumentary | 18 | 1777430 | 1.18% | 6.3E-06 | 15544 | 3.23% | 0.025% | 13036 | 5.05% | 0.032% | 2508 | 1.12% | 0.015% |
| Rosacea | ROSACEA | Integumentary | 19 | 1600896 | 1.06% | 6.0E-06 | 15983 | 3.32% | 0.026% | 13076 | 5.07% | 0.032% | 2907 | 1.30% | 0.016% |
| Seborrheic Dermatitis | SEBDERM | Integumentary | 0 | 1693162 | 1.12% | 6.1E-06 | 13907 | 2.89% | 0.024% | 9300 | 3.60% | 0.027% | 4607 | 2.06% | 0.020% |
| Electrolyte Acid-Base Balance Disorder | ELECTROBAL | Metabolic | 21 | 4063752 | 2.69% | 9.4E-06 | 27615 | 5.73% | 0.033% | 22043 | 8.54% | 0.040% | 5572 | 2.49% | 0.022% |
| Gout Related Crystal Arthropathies | GOUT | Metabolic | 33 | 1782424 | 1.18% | 6.3E-06 | 12483 | 2.59% | 0.023% | 11795 | 4.57% | 0.030% | 688 | 0.31% | 0.008% |
| Malnutrition | MALNUTR | Metabolic | 12 | 535412 | 0.35% | 3.5E-06 | 3220 | 0.67% | 0.012% | 2132 | 0.83% | 0.013% | 1088 | 0.49% | 0.010% |
| Metabolic Syndrome X | METAX | Metabolic | 18 | 708817 | 0.47% | 4.0E-06 | 6020 | 1.25% | 0.016% | 4925 | 1.91% | 0.020% | 1095 | 0.49% | 0.010% |
| Mixed Hyperlipidemia | HYPLIPD | Metabolic | 32 | 5624619 | 3.72% | 1.1E-05 | 41601 | 8.64% | 0.040% | 38811 | 15.04% | 0.052% | 2790 | 1.25% | 0.016% |
| Obesity | OBESE | Metabolic | 15 | 5949861 | 3.94% | 1.1E-05 | 34522 | 7.17% | 0.037% | 27009 | 10.47% | 0.044% | 7513 | 3.36% | 0.026% |
| Overweight | OVERWEIGH | Metabolic | 9 | 862453 | 0.57% | 4.4E-06 | 4761 | 0.99% | 0.014% | 3668 | 1.42% | 0.017% | 1093 | 0.49% | 0.010% |
| Vitamin Deficiency | VITAMIND | Metabolic | 25 | 5716437 | 3.78% | 1.1E-05 | 40787 | 8.47% | 0.040% | 35293 | 13.68% | 0.050% | 5494 | 2.46% | 0.022% |
| Abnormal Spine Curvature | SPINCURV | Musculoskeletal | 10 | 1575373 | 1.04% | 5.9E-06 | 10669 | 2.22% | 0.021% | 5375 | 2.08% | 0.021% | 5294 | 2.37% | 0.022% |
| General Osteochondropathy | OSTCHON | Musculoskeletal | 8 | 527122 | 0.35% | 3.4E-06 | 2641 | 0.55% | 0.011% | 1339 | 0.52% | 0.010% | 1302 | 0.58% | 0.011% |
| General Spondylosis Spine Disorder | SPONSPIN | Musculoskeletal | 21 | 12240293 | 8.10% | 1.6E-05 | 91397 | 18.98% | 0.056% | 76732 | 29.74% | 0.066% | 14665 | 6.56% | 0.036% |
| Hernia | HERNIA | Musculoskeletal | 14 | 3967915 | 2.63% | 9.3E-06 | 31371 | 6.51% | 0.036% | 27264 | 10.57% | 0.044% | 4107 | 1.84% | 0.019% |
| Jaw Disease | JAWD | Musculoskeletal | 14 | 1141076 | 0.76% | 5.0E-06 | 12533 | 2.60% | 0.023% | 7132 | 2.76% | 0.024% | 5401 | 2.41% | 0.022% |
| Muscle Ligament Disorder | MUSCLIGM | Musculoskeletal | 15 | 18273947 | 12.09% | 1.9E-05 | 142750 | 29.64% | 0.066% | 1E+05 | 42.05% | 0.071% | 34279 | 15.33% | 0.052% |
| Spinal Stenosis | SPINSTE | Musculoskeletal | 33 | 2635810 | 1.74% | 7.6E-06 | 19395 | 4.03% | 0.028% | 18070 | 7.00% | 0.037% | 1325 | 0.59% | 0.011% |
| Synovium Tendon Bursa Disorder | SYNTEND | Musculoskeletal | 15 | 5882310 | 3.89% | 1.1E-05 | 56901 | 11.81% | 0.047% | 45019 | 17.45% | 0.055% | 11882 | 5.31% | 0.032% |
| Benign Bone Connective Tissue Neoplasm | BENBONCON | Neoplastic Proce | 12 | 770259 | 0.51% | 4.2E-06 | 7908 | 1.64% | 0.018% | 6196 | 2.40% | 0.022% | 1712 | 0.77% | 0.013% |
| Benign Colon Neoplasm | BENCOLON | Neoplastic Proce | 40 | 5932996 | 3.93% | 1.1E-05 | 67588 | 14.03% | 0.050% | 66012 | 25.59% | 0.063% | 1576 | 0.70% | 0.012% |
| Benign Digestive Neoplasm | BENDIG | Neoplastic Proce | 28 | 933173 | 0.62% | 4.6E-06 | 9363 | 1.94% | 0.020% | 8328 | 3.23% | 0.025% | 1035 | 0.46% | 0.010% |
| Benign Skin Neoplasm | BENSKIN | Neoplastic Proce | 13 | 10148339 | 6.72% | 1.5E-05 | 98600 | 20.47% | 0.058% | 68690 | 26.63% | 0.064% | 29910 | 13.37% | 0.049% |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Benign Uterine Neoplasm | BENUTER | Neoplastic Proce | 29 | 2646498 | 1.75% | 7.7E-06 | 22596 | 4.69% | 0.030% | 21164 | 8.20% | 0.040% | 1432 | 0.64% | 0.011% |
| Cervical Cancer | CERVCAN | Neoplastic Proce | 19 | 1309702 | 0.87% | 5.4E-06 | 16455 | 3.42% | 0.026% | 3861 | 1.50% | 0.017% | 12594 | 5.63% | 0.033% |
| Hemangioma Lymphangioma | HEMLYM | Neoplastic Proce | 5 | 1572112 | 1.04% | 5.9E-06 | 14494 | 3.01% | 0.025% | 12771 | 4.95% | 0.031% | 1723 | 0.77% | 0.013% |
| Lipoma | LIPOMA | Neoplastic Proce | 22 | 1137779 | 0.75% | 5.0E-06 | 11342 | 2.35% | 0.022% | 9696 | 3.76% | 0.027% | 1646 | 0.74% | 0.012% |
| Non-Melanoma Skin Cancer | NONMELANO | Neoplastic Proce | 18 | 8199192 | 5.43% | 1.3E-05 | 74838 | 15.54% | 0.052% | 59277 | 22.98% | 0.061% | 15561 | 6.96% | 0.037% |
| ADHD | ADHD | Neuropsychiatric | 6 | 3343232 | 2.21% | 8.6E-06 | 25141 | 5.22% | 0.032% | 4343 | 1.68% | 0.019% | 20798 | 9.30% | 0.042% |
| Adjustment Disorder | ADJM | Neuropsychiatric | 8 | 4473730 | 2.96% | 9.9E-06 | 36570 | 7.59% | 0.038% | 20527 | 7.96% | 0.039% | 16043 | 7.17% | 0.037% |
| Anxiety Phobic Disorder | GAD | Neuropsychiatric | 13 | 9519463 | 6.30% | 1.4E-05 | 68428 | 14.21% | 0.050% | 36980 | 14.33% | 0.050% | 31448 | 14.06% | 0.050% |
| Autism | ASD | Neuropsychiatric | 2 | 166877 | 0.11% | 1.9E-06 | 1263 | 0.26% | 0.007% | 42 | 0.02% | 0.002% | 1221 | 0.55% | 0.011% |
| Bipolar Disorder | BIP | Neuropsychiatric | 14 | 1282267 | 0.85% | 5.4E-06 | 12215 | 2.54% | 0.023% | 4508 | 1.75% | 0.019% | 7707 | 3.45% | 0.026% |
| Depression | MDD | Neuropsychiatric | 15 | 5732901 | 3.79% | 1.1E-05 | 49650 | 10.31% | 0.044% | 26466 | 10.26% | 0.044% | 23184 | 10.36% | 0.044% |
| Eating Disorder | EATD | Neuropsychiatric | 1 | 239351 | 0.16% | 2.3E-06 | 2708 | 0.56% | 0.011% | 678 | 0.26% | 0.007% | 2030 | 0.91% | 0.014% |
| Migraine | MIGRAINE | Neuropsychiatric | 14 | 4455036 | 2.95% | 9.9E-06 | 35091 | 7.29% | 0.037% | 19967 | 7.74% | 0.039% | 15124 | 6.76% | 0.036% |
| Mood Disorder | MOODD | Neuropsychiatric | 10 | 921174 | 0.61% | 4.5E-06 | 7940 | 1.65% | 0.018% | 2970 | 1.15% | 0.015% | 4970 | 2.22% | 0.021% |
| OCD | OCD | Neuropsychiatric | 9 | 338155 | 0.22% | 2.8E-06 | 3679 | 0.76% | 0.013% | 940 | 0.36% | 0.009% | 2739 | 1.22% | 0.016% |
| PTSD | PTSD | Neuropsychiatric | 12 | 490913 | 0.32% | 3.3E-06 | 3594 | 0.75% | 0.012% | 1848 | 0.72% | 0.012% | 1746 | 0.78% | 0.013% |
| Personality Disorder | PERSON | Neuropsychiatric | 12 | 234341 | 0.16% | 2.3E-06 | 2565 | 0.53% | 0.010% | 876 | 0.34% | 0.008% | 1689 | 0.76% | 0.012% |
| Schizophrenia Related Psychosis | SCHZ | Neuropsychiatric | 16 | 831426 | 0.55% | 4.3E-06 | 6476 | 1.34% | 0.017% | 2759 | 1.07% | 0.015% | 3717 | 1.66% | 0.018% |
| Sleep Disorder | SLEEPD | Neuropsychiatric | 9 | 1088261 | 0.72% | 4.9E-06 | 8670 | 1.80% | 0.019% | 6492 | 2.52% | 0.023% | 2178 | 0.97% | 0.014% |
| Substance Abuse | SUBABUS | Neuropsychiatric | 18 | 6759671 | 4.47% | 1.2E-05 | 47048 | 9.77% | 0.043% | 24798 | 9.61% | 0.042% | 22250 | 9.95% | 0.043% |
| Cataract | CATARACT | Ophthalmologica | 48 | 6736189 | 4.46% | 1.2E-05 | 36763 | 7.63% | 0.038% | 35772 | 13.87% | 0.050% | 991 | 0.44% | 0.010% |
| Conjunctival Disorders | CONJUN | Ophthalmologica | 13 | 1420548 | 0.94% | 5.6E-06 | 12757 | 2.65% | 0.023% | 10614 | 4.11% | 0.029% | 2143 | 0.96% | 0.014% |
| Corneal Disorders | CORNEAL | Ophthalmologica | 16 | 2242583 | 1.48% | 7.1E-06 | 21465 | 4.46% | 0.030% | 13381 | 5.19% | 0.032% | 8084 | 3.61% | 0.027% |
| Glaucoma | GLAUCOMA | Ophthalmologica | 29 | 4363946 | 2.89% | 9.8E-06 | 27955 | 5.80% | 0.034% | 25044 | 9.71% | 0.043% | 2911 | 1.30% | 0.016% |
| Optic Neuritis Neuropathy | OPTNEURO | Ophthalmologica | 13 | 491822 | 0.33% | 3.3E-06 | 3837 | 0.80% | 0.013% | 2854 | 1.11% | 0.015% | 983 | 0.44% | 0.010% |
| Refraction Accomodation Disorders | REFRACC | Ophthalmologica | 9 | 465344 | 0.31% | 3.2E-06 | 3532 | 0.73% | 0.012% | 2553 | 0.99% | 0.014% | 979 | 0.44% | 0.010% |
| Vitreous Body Disorder | VTRBODY | Ophthalmologica | 29 | 2809479 | 1.86% | 7.9E-06 | 23192 | 4.82% | 0.031% | 21122 | 8.19% | 0.040% | 2070 | 0.93% | 0.014% |
| Eustachian Tube Disorder | EUSTUBE | Otic | 1 | 3050373 | 2.02% | 8.2E-06 | 22506 | 4.67% | 0.030% | 15435 | 5.98% | 0.034% | 7071 | 3.16% | 0.025% |
| External Ear Disorders | EXTEAR | Otic | 2 | 4675491 | 3.09% | 1.0E-05 | 33771 | 7.01% | 0.037% | 22820 | 8.85% | 0.041% | 10951 | 4.90% | 0.031% |
| General Ear Disorder | GENEAR | Otic | 1 | 4058482 | 2.69% | 9.4E-06 | 23953 | 4.97% | 0.031% | 14867 | 5.76% | 0.034% | 9086 | 4.06% | 0.028% |
| Hearing Loss | HEARING | Otic | 3 | 4844481 | 3.21% | 1.0E-05 | 37084 | 7.70% | 0.038% | 30677 | 11.89% | 0.047% | 6407 | 2.86% | 0.024% |
| Tympanic Membrane Disorders | TYMPMEM | Otic | 2 | 540569 | 0.36% | 3.5E-06 | 3470 | 0.72% | 0.012% | 1703 | 0.66% | 0.012% | 1767 | 0.79% | 0.013% |
| Cranial Nerve Disorder | CRANNERV | PNS | 4 | 1055892 | 0.70% | 4.9E-06 | 8249 | 1.71% | 0.019% | 6266 | 2.43% | 0.022% | 1983 | 0.89% | 0.014% |
| Nerve Root Plexus Disorders | NERVROOT | PNS | 19 | 593366 | 0.39% | 3.6E-06 | 5112 | 1.06% | 0.015% | 3997 | 1.55% | 0.018% | 1115 | 0.50% | 0.010% |
| Peripheral Nerve Disorder | PERINERV | PNS | 27 | 4225977 | 2.80% | 9.6E-06 | 37149 | 7.71% | 0.038% | 32469 | 12.59% | 0.048% | 4680 | 2.09% | 0.021% |
| Breast Disorder | BREAST | Reproductive | 20 | 6715958 | 4.44% | 1.2E-05 | 56919 | 11.82% | 0.047% | 46449 | 18.01% | 0.055% | 10470 | 4.68% | 0.030% |
| Disease of the Female Reproductive Organs | FEMALEREP | Reproductive | 16 | 15932733 | 10.54% | 1.8E-05 | 118069 | 24.51% | 0.062% | 69025 | 26.76% | 0.064% | 49044 | 21.93% | 0.060% |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Deviated Nasal Septum | DEVNASAL | Respiratory | 9 | 981185 | 0.65% | 4.7E-06 | 9010 | 1.87% | 0.020% | 5540 | 2.15% | 0.021% | 3470 | 1.55% | 0.018% |
| Emphysema COPD | COPD | Respiratory | 15 | 2212557 | 1.46% | 7.0E-06 | 11195 | 2.32% | 0.022% | 9331 | 3.62% | 0.027% | 1864 | 0.83% | 0.013% |
| Sleep Apnea | SLEEPAPN | Respiratory | 20 | 3585952 | 2.37% | 8.9E-06 | 29760 | 6.18% | 0.035% | 26948 | 10.45% | 0.044% | 2812 | 1.26% | 0.016% |
| Bladder Disorder | BLADDER | Urinary | 14 | 1178825 | 0.78% | 5.1E-06 | 9510 | 1.97% | 0.020% | 7696 | 2.98% | 0.025% | 1814 | 0.81% | 0.013% |
| Urinary Calculus | URINARY | Urinary | 22 | 2758952 | 1.83% | 7.8E-06 | 23853 | 4.95% | 0.031% | 18941 | 7.34% | 0.038% | 4912 | 2.20% | 0.021% |

**Table A.2:** Heritability and preventability estimates and standard deviations for 149 studied diseases

**Supplementary Table 2. Heritability and preventability estimates and standard deviations for 149 studied diseases**

| Disease | Acronym | Selected Model | $h^2$ | $h^2$ Before Adjustment | $h^2$ SD | Couple | Couple SD | Sibling | Sibling SD | Unique | Unique SD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Brain Damage | BRAIND | GC | 0.226 | 0.214 | 0.031 | 0.091 | 0.029 | - | - | 0.695 | 0.039 |
| Disorder of the Vestibular System | DVS | GCS | 0.295 | 0.279 | 0.024 | 0.104 | 0.014 | 0.103 | 0.046 | 0.514 | 0.050 |
| Epilepsy Related Disorders | EPILEPSY | GC | 0.541 | 0.512 | 0.019 | 0.249 | 0.020 | - | - | 0.239 | 0.026 |
| Extrapyramidal Abnormal Movement Disorders | EAM | GCS | 0.262 | 0.248 | 0.025 | 0.117 | 0.018 | 0.012 | 0.014 | 0.624 | 0.034 |
| Ocular Musculoskeletal Disorder | OMD | GCS | 0.357 | 0.338 | 0.021 | 0.211 | 0.015 | 0.200 | 0.024 | 0.251 | 0.024 |
| Aneurysm | ANEURYSM | GCS | 0.695 | 0.658 | 0.018 | 0.212 | 0.017 | 0.000 | 0.000 | 0.129 | 0.010 |
| Atherosclerosis | ATHRSLS | GCS | 0.406 | 0.384 | 0.022 | 0.205 | 0.007 | 0.196 | 0.028 | 0.214 | 0.020 |
| Cardiac Dysrhythmia | CARDYS | GCS | 0.240 | 0.227 | 0.011 | 0.144 | 0.006 | 0.018 | 0.014 | 0.610 | 0.018 |
| Cardiomyopathy | CARDMYO | GCS | 0.604 | 0.572 | 0.022 | 0.210 | 0.018 | 0.000 | 0.000 | 0.218 | 0.007 |
| Cerebrovascular Disease | CEREBVAS | GCS | 0.217 | 0.205 | 0.021 | 0.161 | 0.009 | 0.125 | 0.046 | 0.509 | 0.048 |
| General Hypertension | HYPTEN | GCS | 0.462 | 0.438 | 0.009 | 0.195 | 0.004 | 0.020 | 0.013 | 0.347 | 0.015 |
| Non-Rheumatic Heart Disease | NONRHHD | GCS | 0.344 | 0.326 | 0.014 | 0.190 | 0.008 | 0.102 | 0.027 | 0.382 | 0.029 |
| Peripheral Vascular Disease | PERIVAS | GCS | 0.253 | 0.239 | 0.022 | 0.152 | 0.010 | 0.128 | 0.042 | 0.481 | 0.041 |
| Cardiac Congenital Anomaly | CARCONG | GCS | 0.393 | 0.372 | 0.028 | 0.226 | 0.020 | 0.216 | 0.034 | 0.186 | 0.014 |
| Oro-Facial Congenital Anomaly | OFCONG | GCS | 0.440 | 0.416 | 0.027 | 0.205 | 0.017 | 0.258 | 0.017 | 0.120 | 0.008 |
| Appendiceal Disease | APPEND | GCS | 0.260 | 0.246 | 0.027 | 0.138 | 0.028 | 0.000 | 0.000 | 0.616 | 0.036 |
| Esophageal Disease | ESOPHD | GC | 0.292 | 0.277 | 0.008 | 0.130 | 0.005 | - | - | 0.593 | 0.009 |
| Functional Digestive Disorder | DIGEST | GCS | 0.203 | 0.192 | 0.009 | 0.119 | 0.007 | 0.019 | 0.013 | 0.670 | 0.016 |
| Gastrointestinal Ulcer | GASTROULC | GC | 0.306 | 0.289 | 0.026 | 0.141 | 0.015 | - | - | 0.570 | 0.030 |
| Rectal Anal Disorder | RECTANAL | GCS | 0.078 | 0.074 | 0.022 | 0.083 | 0.014 | 0.001 | 0.001 | 0.842 | 0.026 |
| Goiter | GOITER | GCS | 0.408 | 0.386 | 0.017 | 0.145 | 0.012 | 0.054 | 0.026 | 0.415 | 0.032 |
| Thyrotoxicosis | THYRODTOX | GCS | 0.405 | 0.384 | 0.023 | 0.226 | 0.016 | 0.134 | 0.031 | 0.256 | 0.024 |
| Type II Diabetes Mellitus | DM2 | GCS | 0.561 | 0.532 | 0.010 | 0.283 | 0.005 | 0.112 | 0.013 | 0.074 | 0.006 |
| Acquired Coagulation Defect | ACQCOAG | GCS | 0.642 | 0.608 | 0.019 | 0.233 | 0.017 | 0.000 | 0.000 | 0.158 | 0.014 |
| Pernicious or B12 Deficiency Anemia | B12DEF | GC | 0.652 | 0.617 | 0.013 | 0.303 | 0.012 | - | - | 0.080 | 0.004 |
| Biliary Tract Disease | BILLARY | GCS | 0.219 | 0.207 | 0.013 | 0.101 | 0.009 | 0.083 | 0.021 | 0.608 | 0.024 |
| Chronic Liver Disease | CHRLIVER | GCS | 0.363 | 0.344 | 0.025 | 0.158 | 0.011 | 0.248 | 0.031 | 0.250 | 0.025 |
| Allergic Rhinitis | ALLGRHIN | GCS | 0.445 | 0.421 | 0.006 | 0.203 | 0.005 | 0.008 | 0.006 | 0.368 | 0.009 |
| Asthma | ASTHMA | GCS | 0.457 | 0.433 | 0.008 | 0.151 | 0.007 | 0.009 | 0.007 | 0.407 | 0.012 |
| Atopic Contact Dermatitis | DERM | GCS | 0.202 | 0.191 | 0.006 | 0.108 | 0.005 | 0.006 | 0.005 | 0.694 | 0.009 |
| Chronic Sinusitis | CHRSINUS | GCS | 0.523 | 0.495 | 0.008 | 0.249 | 0.006 | 0.036 | 0.010 | 0.220 | 0.015 |
| Chronic Tonsillitis Adenoiditis | TONSADEN | GCS | 0.330 | 0.312 | 0.021 | 0.213 | 0.034 | 0.007 | 0.008 | 0.468 | 0.039 |
| Crohns | CROHNS | GCS | 0.574 | 0.544 | 0.028 | 0.181 | 0.027 | 0.003 | 0.004 | 0.272 | 0.029 |
| Cystitis Urethritis | CYSURETH | GCS | 0.313 | 0.296 | 0.015 | 0.173 | 0.014 | 0.003 | 0.003 | 0.528 | 0.020 |
| Esophagitis | ESOPHI | GCS | 0.364 | 0.345 | 0.019 | 0.201 | 0.011 | 0.117 | 0.030 | 0.337 | 0.033 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Eye Inflammation | EYEINFM | GCS | 0.292 | 0.276 | 0.009 | 0.148 | 0.007 | 0.061 | 0.010 | 0.515 | 0.012 |
| Fasciitis | FASCI | GCS | 0.350 | 0.331 | 0.014 | 0.132 | 0.007 | 0.089 | 0.026 | 0.448 | 0.028 |
| Food Allergy | FOODALL | GC | 0.599 | 0.567 | 0.021 | 0.315 | 0.021 | - | 0.000 | 0.118 | 0.007 |
| Gastritis Duodenitis | GASDUO | GCS | 0.288 | 0.273 | 0.010 | 0.188 | 0.007 | 0.002 | 0.003 | 0.537 | 0.013 |
| IBS | IBS | GCS | 0.287 | 0.272 | 0.015 | 0.164 | 0.013 | 0.015 | 0.014 | 0.548 | 0.021 |
| Inflammatory Spondylopathies | INFMSPON | GC | 0.646 | 0.612 | 0.016 | 0.269 | 0.015 | - | - | 0.119 | 0.009 |
| Lupus Erythematosus | LUPUS | GC | 0.660 | 0.625 | 0.027 | 0.252 | 0.027 | - | - | 0.123 | 0.007 |
| Lymphatic Disorder | LYMPHD | GCS | 0.215 | 0.204 | 0.027 | 0.149 | 0.021 | 0.144 | 0.028 | 0.504 | 0.038 |
| Osteoarthritis | OSTARTH | GCS | 0.256 | 0.242 | 0.012 | 0.134 | 0.005 | 0.009 | 0.012 | 0.615 | 0.017 |
| Osteomyelitis | OSTMYEL | GCS | 0.220 | 0.208 | 0.023 | 0.154 | 0.019 | 0.000 | 0.000 | 0.638 | 0.030 |
| Pleuritis | PLEUI | GCS | 0.184 | 0.174 | 0.024 | 0.115 | 0.015 | 0.002 | 0.003 | 0.709 | 0.029 |
| Psoriasis Related Disorders | PSORIAS | GCS | 0.304 | 0.288 | 0.019 | 0.082 | 0.016 | 0.009 | 0.011 | 0.621 | 0.028 |
| Rheumatic Heart Disease | RHHD | GC | 0.554 | 0.524 | 0.017 | 0.288 | 0.013 | - | - | 0.188 | 0.016 |
| Rheumatoid Arthritis Related Conditions | RA | GCS | 0.486 | 0.460 | 0.027 | 0.226 | 0.014 | 0.068 | 0.031 | 0.247 | 0.022 |
| Secondary Arthropathy | 2NDARTHR | GCS | 0.629 | 0.596 | 0.017 | 0.241 | 0.016 | 0.006 | 0.007 | 0.158 | 0.014 |
| Thyroiditis | THYRODI | GC | 0.635 | 0.601 | 0.015 | 0.261 | 0.014 | - | - | 0.138 | 0.011 |
| Type I Diabetes Mellitus | DM1 | GCS | 0.531 | 0.503 | 0.020 | 0.244 | 0.012 | 0.127 | 0.023 | 0.126 | 0.011 |
| Ulcerative Colitis | ULCCOL | GCS | 0.471 | 0.446 | 0.033 | 0.160 | 0.024 | 0.136 | 0.039 | 0.258 | 0.019 |
| Upper Respiratory Inflammation | UPRESP | GCS | 0.287 | 0.272 | 0.011 | 0.148 | 0.009 | 0.006 | 0.007 | 0.574 | 0.015 |
| Arthropod-Borne Diseases | ARTHROPOD | GCS | 0.639 | 0.605 | 0.023 | 0.307 | 0.022 | 0.000 | 0.000 | 0.088 | 0.006 |
| Cellulitis | CELLUI | GCS | 0.226 | 0.214 | 0.007 | 0.130 | 0.006 | 0.000 | 0.000 | 0.656 | 0.009 |
| Chronic Upper Respiratory Infection | CHRUPRESP | GCS | 0.464 | 0.439 | 0.013 | 0.229 | 0.010 | 0.064 | 0.017 | 0.268 | 0.020 |
| Ear Infection | EARINF | GCS | 0.244 | 0.231 | 0.007 | 0.131 | 0.006 | 0.011 | 0.008 | 0.626 | 0.011 |
| Eye Infection | EYEINF | GCS | 0.200 | 0.189 | 0.009 | 0.114 | 0.007 | 0.031 | 0.010 | 0.665 | 0.013 |
| Fungal Infection | FUNGINF | GCS | 0.211 | 0.200 | 0.007 | 0.144 | 0.006 | 0.000 | 0.000 | 0.656 | 0.009 |
| Gastrointestinal Infection | GASTROINF | GCS | 0.295 | 0.279 | 0.016 | 0.188 | 0.014 | 0.066 | 0.021 | 0.467 | 0.029 |
| General Bacterial Infection | BACTINF | GCS | 0.409 | 0.387 | 0.032 | 0.240 | 0.024 | 0.140 | 0.035 | 0.232 | 0.022 |
| General Viral Infection | VIRINF | GC | 0.267 | 0.253 | 0.010 | 0.206 | 0.010 | - | - | 0.541 | 0.014 |
| Herpes Simplex | HERPSX | GCS | 0.188 | 0.178 | 0.018 | 0.208 | 0.016 | 0.000 | 0.000 | 0.614 | 0.025 |
| Herpes Zoster | HERPZOS | GCS | 0.127 | 0.120 | 0.023 | 0.097 | 0.013 | 0.000 | 0.000 | 0.783 | 0.027 |
| Infestation | INFESTA | GCS | 0.463 | 0.438 | 0.019 | 0.457 | 0.019 | 0.000 | 0.000 | 0.105 | 0.007 |
| Influenza | FLU | GCS | 0.412 | 0.390 | 0.013 | 0.291 | 0.011 | 0.054 | 0.015 | 0.265 | 0.020 |
| Kidney Infection | KIDNEYINF | GC | 0.221 | 0.209 | 0.028 | 0.157 | 0.026 | - | - | 0.633 | 0.040 |
| Pneumonia | PNEUMO | GCS | 0.391 | 0.370 | 0.030 | 0.226 | 0.019 | 0.143 | 0.037 | 0.261 | 0.030 |
| STI | STI | GC | 0.519 | 0.491 | 0.022 | 0.365 | 0.021 | 0.000 | 0.000 | 0.144 | 0.013 |
| Septicemia | SEPSIS | GCS | 0.391 | 0.371 | 0.033 | 0.186 | 0.020 | 0.287 | 0.035 | 0.157 | 0.012 |
| Skin Infection | SKININF | GCS | 0.243 | 0.230 | 0.016 | 0.138 | 0.015 | 0.004 | 0.005 | 0.629 | 0.022 |
| UTI | UTI | GCS | 0.227 | 0.215 | 0.007 | 0.139 | 0.005 | 0.022 | 0.010 | 0.624 | 0.011 |
| Viral Hepatitis B | HEPAB | GC | 0.540 | 0.512 | 0.040 | 0.329 | 0.036 | - | - | 0.159 | 0.017 |
| Viral Warts HPV | HPV | GCS | 0.289 | 0.274 | 0.009 | 0.176 | 0.008 | 0.029 | 0.009 | 0.521 | 0.012 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Acne | ACNE | GCS | 0.501 | 0.474 | 0.010 | 0.195 | 0.015 | 0.080 | 0.008 | 0.250 | 0.018 |
| Alopecia | ALOPE | GCS | 0.321 | 0.304 | 0.023 | 0.122 | 0.025 | 0.047 | 0.025 | 0.526 | 0.037 |
| Carbuncle Furuncle | CARBFUR | GC | 0.424 | 0.402 | 0.024 | 0.288 | 0.019 | 0.000 | 0.000 | 0.310 | 0.029 |
| Chronic Pruritus | CHRPRUR | GCS | 0.227 | 0.215 | 0.023 | 0.150 | 0.015 | 0.051 | 0.030 | 0.584 | 0.035 |
| Dyschromia | DYSCHROM | GCS | 0.430 | 0.407 | 0.014 | 0.330 | 0.007 | 0.114 | 0.016 | 0.148 | 0.010 |
| Hair Related Disease | HAIRD | GC | 0.238 | 0.225 | 0.013 | 0.102 | 0.015 | - | - | 0.672 | 0.021 |
| Keratosis | KERATO | GCS | 0.344 | 0.326 | 0.015 | 0.320 | 0.004 | 0.092 | 0.025 | 0.262 | 0.023 |
| Nail Disease | NAILD | GC | 0.325 | 0.308 | 0.014 | 0.157 | 0.010 | - | - | 0.535 | 0.018 |
| Oral Soft Tissue Disease | ORALSOFT | GCS | 0.113 | 0.107 | 0.020 | 0.091 | 0.016 | 0.060 | 0.027 | 0.742 | 0.031 |
| Photodermatitis | PHOTODERM | GCS | 0.516 | 0.488 | 0.008 | 0.462 | 0.007 | 0.000 | 0.000 | 0.049 | 0.002 |
| Rosacea | ROSACEA | GCS | 0.397 | 0.376 | 0.019 | 0.188 | 0.011 | 0.079 | 0.031 | 0.357 | 0.029 |
| Seborrheic Dermatitis | SEBDERM | GCS | 0.372 | 0.352 | 0.018 | 0.140 | 0.013 | 0.044 | 0.026 | 0.465 | 0.033 |
| Electrolyte Acid-Base Balance Disorder | ELECTROBAL | GCS | 0.288 | 0.273 | 0.014 | 0.155 | 0.009 | 0.048 | 0.027 | 0.524 | 0.030 |
| Gout Related Crystal Arthropathies | GOUT | GC | 0.490 | 0.464 | 0.023 | 0.175 | 0.014 | - | - | 0.361 | 0.027 |
| Malnutrition | MALNUTR | GC | 0.539 | 0.511 | 0.023 | 0.318 | 0.021 | - | - | 0.171 | 0.014 |
| Metabolic Syndrome X | METAX | GC | 0.630 | 0.596 | 0.014 | 0.332 | 0.013 | - | - | 0.071 | 0.004 |
| Mixed Hyperlipidemia | HYPLIPD | GC | 0.610 | 0.577 | 0.005 | 0.387 | 0.005 | - | - | 0.036 | 0.002 |
| Obesity | OBESE | GCS | 0.640 | 0.606 | 0.007 | 0.293 | 0.006 | 0.000 | 0.000 | 0.101 | 0.007 |
| Overweight | OVERWEIGH | GCS | 0.576 | 0.546 | 0.016 | 0.329 | 0.015 | 0.000 | 0.000 | 0.125 | 0.006 |
| Vitamin Deficiency | VITAMIND | GCS | 0.629 | 0.596 | 0.007 | 0.323 | 0.005 | 0.000 | 0.000 | 0.081 | 0.006 |
| Abnormal Spine Curvature | SPINCURV | GCS | 0.328 | 0.311 | 0.018 | 0.194 | 0.019 | 0.005 | 0.006 | 0.491 | 0.026 |
| General Osteochondropathy | OSTCHON | GCS | 0.491 | 0.465 | 0.029 | 0.243 | 0.031 | 0.000 | 0.000 | 0.292 | 0.029 |
| General Spondylosis Spine Disorder | SPONSPIN | GCS | 0.325 | 0.308 | 0.008 | 0.154 | 0.005 | 0.037 | 0.012 | 0.502 | 0.014 |
| Hernia | HERNIA | GCS | 0.173 | 0.164 | 0.014 | 0.084 | 0.008 | 0.003 | 0.004 | 0.749 | 0.017 |
| Jaw Disease | JAWD | GC | 0.205 | 0.194 | 0.018 | 0.086 | 0.019 | - | - | 0.720 | 0.027 |
| Muscle Ligament Disorder | MUSCLIGM | GCS | 0.268 | 0.254 | 0.006 | 0.148 | 0.004 | 0.001 | 0.001 | 0.597 | 0.008 |
| Spinal Stenosis | SPINSTE | GCS | 0.343 | 0.324 | 0.023 | 0.121 | 0.009 | 0.181 | 0.034 | 0.374 | 0.030 |
| Synovium Tendon Bursa Disorder | SYNTEND | GCS | 0.180 | 0.171 | 0.009 | 0.111 | 0.006 | 0.030 | 0.014 | 0.688 | 0.016 |
| Benign Bone Connective Tissue Neoplasm | BENBONCON | GC | 0.218 | 0.206 | 0.027 | 0.109 | 0.017 | - | - | 0.685 | 0.031 |
| Benign Colon Neoplasm | BENCOLON | GCS | 0.173 | 0.164 | 0.019 | 0.171 | 0.005 | 0.217 | 0.040 | 0.448 | 0.042 |
| Benign Digestive Neoplasm | BENDIG | GC | 0.506 | 0.479 | 0.031 | 0.186 | 0.015 | - | - | 0.335 | 0.036 |
| Benign Skin Neoplasm | BENSKIN | GCS | 0.547 | 0.518 | 0.007 | 0.312 | 0.004 | 0.036 | 0.008 | 0.135 | 0.009 |
| Benign Uterine Neoplasm | BENUTER | GCS | 0.162 | 0.154 | 0.029 | 0.352 | 0.035 | 0.262 | 0.044 | 0.232 | 0.018 |
| Cervical Cancer | CERVCAN | GCS | 0.238 | 0.225 | 0.021 | 0.345 | 0.036 | 0.060 | 0.017 | 0.370 | 0.037 |
| Hemangioma Lymphangioma | HEMLYM | GCS | 0.377 | 0.357 | 0.023 | 0.279 | 0.010 | 0.000 | 0.000 | 0.364 | 0.026 |
| Lipoma | LIPOMA | GCS | 0.038 | 0.036 | 0.027 | 0.071 | 0.014 | 0.137 | 0.048 | 0.756 | 0.051 |
| Non-Melanoma Skin Cancer | NONMELANO | GCS | 0.520 | 0.493 | 0.008 | 0.291 | 0.005 | 0.033 | 0.012 | 0.183 | 0.016 |
| ADHD | ADHD | GC | 0.763 | 0.723 | 0.009 | 0.200 | 0.009 | - | - | 0.077 | 0.007 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Adjustment Disorder | ADJM | GCS | 0.561 | 0.531 | 0.006 | 0.422 | 0.006 | 0.000 | 0.000 | 0.047 | 0.004 |
| Anxiety Phobic Disorder | GAD | GCS | 0.432 | 0.409 | 0.007 | 0.240 | 0.006 | 0.001 | 0.002 | 0.349 | 0.010 |
| Autism | ASD | GCS | 0.924 | 0.875 | 0.006 | 0.000 | 0.000 | 0.000 | 0.000 | 0.125 | 0.006 |
| Bipolar Disorder | BIP | GCS | 0.676 | 0.640 | 0.011 | 0.269 | 0.011 | 0.000 | 0.000 | 0.091 | 0.007 |
| Depression | MDD | GC | 0.579 | 0.548 | 0.006 | 0.362 | 0.006 | 0.001 | 0.001 | 0.089 | 0.008 |
| Eating Disorder | EATD | GC | 0.569 | 0.539 | 0.031 | 0.317 | 0.034 | - | - | 0.144 | 0.012 |
| Migraine | MIGRAINE | GCS | 0.374 | 0.354 | 0.010 | 0.108 | 0.010 | 0.003 | 0.004 | 0.534 | 0.015 |
| Mood Disorder | MOODD | GCS | 0.521 | 0.493 | 0.021 | 0.302 | 0.018 | 0.028 | 0.019 | 0.176 | 0.014 |
| OCD | OCD | GC | 0.657 | 0.622 | 0.023 | 0.267 | 0.023 | - | - | 0.111 | 0.007 |
| PTSD | PTSD | GC | 0.577 | 0.546 | 0.021 | 0.325 | 0.021 | - | - | 0.128 | 0.009 |
| Personality Disorder | PERSON | GCS | 0.520 | 0.492 | 0.028 | 0.394 | 0.028 | 0.000 | 0.000 | 0.114 | 0.007 |
| Schizophrenia Related Psychosis | SCHZ | GCS | 0.562 | 0.532 | 0.022 | 0.253 | 0.020 | 0.025 | 0.020 | 0.189 | 0.019 |
| Sleep Disorder | SLEEPD | GCS | 0.380 | 0.360 | 0.025 | 0.223 | 0.015 | 0.148 | 0.031 | 0.269 | 0.028 |
| Substance Abuse | SUBABUS | GCS | 0.422 | 0.399 | 0.010 | 0.341 | 0.007 | 0.026 | 0.011 | 0.234 | 0.016 |
| Cataract | CATARACT | GCS | 0.310 | 0.293 | 0.024 | 0.332 | 0.006 | 0.203 | 0.028 | 0.172 | 0.015 |
| Conjunctival Disorders | CONJUN | GCS | 0.346 | 0.328 | 0.022 | 0.150 | 0.012 | 0.164 | 0.030 | 0.358 | 0.034 |
| Corneal Disorders | CORNEAL | GCS | 0.317 | 0.300 | 0.014 | 0.182 | 0.011 | 0.078 | 0.017 | 0.439 | 0.022 |
| Glaucoma | GLAUCOMA | GCS | 0.698 | 0.661 | 0.008 | 0.215 | 0.006 | 0.000 | 0.000 | 0.124 | 0.007 |
| Optic Neuritis Neuropathy | OPTNEURO | GC | 0.593 | 0.562 | 0.017 | 0.323 | 0.017 | - | - | 0.115 | 0.007 |
| Refraction Accomodation Disorders | REFRACC | GC | 0.611 | 0.579 | 0.015 | 0.384 | 0.015 | - | - | 0.037 | 0.001 |
| Vitreous Body Disorder | VTRBODY | GCS | 0.484 | 0.458 | 0.015 | 0.206 | 0.008 | 0.218 | 0.019 | 0.119 | 0.010 |
| Eustachian Tube Disorder | EUSTUBE | GCS | 0.274 | 0.260 | 0.014 | 0.126 | 0.010 | 0.003 | 0.003 | 0.612 | 0.018 |
| External Ear Disorders | EXTEAR | GCS | 0.272 | 0.257 | 0.012 | 0.106 | 0.008 | 0.054 | 0.016 | 0.582 | 0.019 |
| General Ear Disorder | GENEAR | GCS | 0.243 | 0.230 | 0.013 | 0.103 | 0.010 | 0.015 | 0.013 | 0.652 | 0.020 |
| Hearing Loss | HEARING | GCS | 0.217 | 0.206 | 0.012 | 0.095 | 0.007 | 0.054 | 0.023 | 0.645 | 0.025 |
| Tympanic Membrane Disorders | TYMPMEM | GCS | 0.298 | 0.283 | 0.040 | 0.131 | 0.040 | 0.000 | 0.000 | 0.586 | 0.057 |
| Cranial Nerve Disorder | CRANNERV | GCS | 0.249 | 0.236 | 0.027 | 0.112 | 0.018 | 0.096 | 0.037 | 0.557 | 0.041 |
| Nerve Root Plexus Disorders | NERVROOT | GCS | 0.656 | 0.621 | 0.014 | 0.314 | 0.014 | 0.000 | 0.000 | 0.065 | 0.004 |
| Peripheral Nerve Disorder | PERINERV | GCS | 0.213 | 0.201 | 0.014 | 0.111 | 0.007 | 0.000 | 0.000 | 0.688 | 0.016 |
| Breast Disorder | BREAST | GCS | 0.166 | 0.157 | 0.010 | 0.091 | 0.012 | 0.001 | 0.001 | 0.752 | 0.015 |
| Disease of the Female Reproductive Organs | FEMALEREP | GCS | 0.235 | 0.223 | 0.009 | 0.138 | 0.018 | 0.080 | 0.011 | 0.560 | 0.022 |
| Deviated Nasal Septum | DEVNASAL | GC | 0.489 | 0.463 | 0.019 | 0.255 | 0.017 | - | - | 0.282 | 0.027 |
| Emphysema COPD | COPD | GCS | 0.534 | 0.506 | 0.019 | 0.281 | 0.011 | 0.122 | 0.021 | 0.091 | 0.006 |
| Sleep Apnea | SLEEPAPN | GCS | 0.474 | 0.449 | 0.015 | 0.209 | 0.008 | 0.000 | 0.000 | 0.343 | 0.017 |
| Bladder Disorder | BLADDER | GC | 0.223 | 0.211 | 0.027 | 0.131 | 0.015 | - | - | 0.657 | 0.031 |
| Urinary Calculus | URINARY | GCS | 0.357 | 0.338 | 0.015 | 0.103 | 0.009 | 0.052 | 0.027 | 0.507 | 0.032 |

**Table A.3: Heritability estimates compared to published twin/family studies. Bold** font indicates the two estimates' 95% confidence intervals do not overlap with each other

| Disease | Family $h^2$ Adjusted | Family $h^2$ SD | Reference $h^2$ | Reference $h^2$ SD | Reference |
|---|---|---|---|---|---|
| Disorder of the Vestibular System | 0.295 | 0.024 | 0.31 | 0.042 | [226] |
| Epilepsy Related Disorders | 0.541 | 0.019 | 0.7 | 0.08 | [151] |
| Aneurysm | 0.695 | 0.018 | 0.77 | 0.045 | [143] |
| Cardiac Dysrhythmia | 0.24 | 0.011 | 0.27 | 0.07 | [226] |
| Cerebrovascular Disease | 0.217 | 0.021 | 0.23 | 0.01 | [199] |
| General Hypertension | 0.462 | 0.009 | 0.37 | 0.09 | [297] |
| Non-Rheumatic Heart Disease | 0.344 | 0.014 | 0.29 | 0.037 | [226] |
| Peripheral Vascular Disease | 0.253 | 0.022 | 0.58 | 0.035 | [301] |
| Appendiceal Disease | 0.26 | 0.027 | 0.2 | 0.1 | [245] |
| Esophageal Disease | 0.292 | 0.008 | 0.256 | 0.044 | [226] |
| Goiter | 0.408 | 0.017 | 0.317 | 0.06 | [227] |
| Type II Diabetes Mellitus | 0.561 | 0.01 | 0.46 | 0.1 | [310] |
| Biliary Tract Disease | 0.219 | 0.013 | 0.42 | 0.1575 | [257] |
| Chronic Liver Disease | 0.363 | 0.025 | 0.52 | 0.11 | [177] |

Continued on next page

**Table A.3 – continued from previous page**

| Disease | Family $h^2$ Adjusted | Family $h^2$ SD | Reference $h^2$ | Reference $h^2$ SD | Reference |
|---|---|---|---|---|---|
| Allergic Rhinitis | 0.445 | 0.006 | 0.551 | 0.07 | [226] |
| Asthma | 0.457 | 0.008 | 0.535 | 0.048 | [226] |
| Atopic Contact Dermatitis | 0.202 | 0.006 | 0.86 | 0.045 | [140] |
| Chronic Sinusitis | 0.523 | 0.008 | 0.33 | 0.1 | [12] |
| Chronic Tonsillitis Adenoiditis | 0.33 | 0.021 | 0.33 | 0.1 | [226] |
| Crohn's Disease | 0.574 | 0.028 | 0.786 | 0.143 | [42] |
| Cystitis Urethritis | 0.313 | 0.015 | 0.399 | 0.072 | [226] |
| Esophagitis | 0.364 | 0.019 | 0.145 | 0.04 | [4] |
| Food Allergy | 0.599 | 0.021 | 0.63 | 0.041 | [226] |
| IBS | 0.287 | 0.015 | 0.217 | 0.043 | [226] |
| Inflammatory Spondylopathies | 0.646 | 0.016 | 0.97 | 0.025 | [32] |
| Lupus Erythematosus | 0.66 | 0.027 | 0.66 | 0.11 | [164] |
| Osteoarthritis | 0.256 | 0.012 | 0.332 | 0.121 | [226] |
| Psoriasis Related Disorders | 0.304 | 0.019 | 0.571 | 0.078 | [226] |
| Rheumatoid Arthritis Related Conditions | 0.486 | 0.027 | 0.53 | 0.065 | [183] |
| Type I Diabetes Mellitus | 0.531 | 0.02 | 0.496 | 0.125 | [226] |

Continued on next page

**Table A.3 – continued from previous page**

| Disease | Family $h^2$ Adjusted | Family $h^2$ SD | Reference $h^2$ | Reference $h^2$ SD | Reference |
|---|---|---|---|---|---|
| Ulcerative Colitis | 0.471 | 0.033 | 0.62 | 0.042 | [289] |
| Ear Infection | 0.244 | 0.007 | 0.49 | 0.05 | [239] |
| Viral Hepatitis B | 0.54 | 0.04 | 0.503 | 0.1185 | [131] |
| Acne | 0.501 | 0.01 | 0.81 | 0.04 | [17] |
| Alopecia | 0.321 | 0.023 | 0.81 | 0.02 | [113] |
| Rosacea | 0.397 | 0.019 | 0.46 | 0.05 | [3] |
| Electrolyte Acid-Base Balance Disorder | 0.288 | 0.014 | 0.319 | 0.026 | [226] |
| Gout Related Crystal Arthropathies | 0.49 | 0.023 | 0.469 | 0.08 | [158] |
| Obesity | 0.64 | 0.007 | 0.631 | 0.013 | [226] |
| Abnormal Spine Curvature | 0.328 | 0.018 | 0.38 | 0.04 | [103] |
| Spinal Stenosis | 0.343 | 0.023 | 0.669 | 0.05 | [19] |
| Cervical Cancer | 0.238 | 0.021 | 0.22 | 0.03 | [54] |
| Non-Melanoma Skin Cancer | 0.52 | 0.008 | 0.43 | 0.08 | [197] |
| ADHD | 0.763 | 0.009 | 0.71 | 0.029 | [206] |
| Anxiety Phobic Disorder | 0.432 | 0.007 | 0.49 | 0.026 | [226] |
| Autism | 0.924 | 0.006 | 0.775 | 0.0675 | [284] |

Continued on next page

**Table A.3 – continued from previous page**

| Disease | Family $h^2$ Adjusted | Family $h^2$ SD | Reference $h^2$ | Reference $h^2$ SD | Reference |
|---|---|---|---|---|---|
| Bipolar Disorder | 0.676 | 0.011 | 0.678 | 0.02 | [226] |
| Depression | 0.579 | 0.006 | 0.402 | 0.016 | [226] |
| Eating Disorder | 0.569 | 0.031 | 0.56 | 0.2 | [34] |
| Migraine | 0.374 | 0.01 | 0.415 | 0.026 | [226] |
| Mood Disorder | 0.521 | 0.021 | 0.609 | 0.086 | [226] |
| OCD | 0.657 | 0.023 | 0.448 | 0.026 | [226] |
| PTSD | 0.577 | 0.021 | 0.46 | 0.08 | [253] |
| Personality Disorder | 0.52 | 0.028 | 0.622 | 0.052 | [226] |
| Schizophrenia Related Psychosis | 0.562 | 0.022 | 0.643 | 0.013 | [226] |
| Sleep Disorder | 0.38 | 0.025 | 0.345 | 0.026 | [226] |
| Substance Abuse | 0.422 | 0.01 | 0.408 | 0.051 | [226] |
| Cataract | 0.31 | 0.024 | 0.356 | 0.07 | [252] |
| Glaucoma | 0.698 | 0.008 | 0.665 | 0.03 | [252] |
| Optic Neuritis Neuropathy | 0.593 | 0.017 | 0.52 | 0.31 | [252] |
| Refraction Accommodation Disorders | 0.611 | 0.015 | 0.58 | 0.13 | [252] |
| Hearing Loss | 0.217 | 0.012 | 0.25 | 0.045 | [194] |

Continued on next page

**Table A.3 – continued from previous page**

| Disease | Family $h^2$ Adjusted | Family $h^2$ SD | Reference $h^2$ | Reference $h^2$ SD | Reference |
|---|---|---|---|---|---|
| Emphysema COPD | 0.534 | 0.019 | 0.63 | 0.08 | [136] |
| Sleep Apnea | 0.474 | 0.015 | 0.37 | 0.04 | [221] |
| Urinary Calculus | 0.357 | 0.015 | 0.52 | 0.05 | [133] |

**Table A.4:** Genetic, environmental and phenotypic correlations of pairwise analysis for 29 complex diseases

| Disease1 | Disease2 | $r_g$ | $r_g$ SD | $r_e$ | $r_e$ SD | $r_p$ | $r_p$ SD |
|---|---|---|---|---|---|---|---|
| ADHD | Adjustment Disorder | 0.358 | 0.013 | 0.192 | 0.020 | 0.291 | 0.005 |
| ADHD | Anxiety Phobic Disorder | 0.447 | 0.012 | 0.226 | 0.017 | 0.334 | 0.004 |
| ADHD | Bipolar Disorder | 0.368 | 0.013 | 0.343 | 0.025 | 0.358 | 0.006 |
| ADHD | Depression | 0.421 | 0.011 | 0.274 | 0.019 | 0.361 | 0.004 |
| ADHD | Migraine | 0.216 | 0.017 | 0.044 | 0.020 | 0.127 | 0.005 |
| ADHD | Mood Disorder | 0.401 | 0.019 | 0.244 | 0.028 | 0.331 | 0.007 |
| ADHD | Schizophrenia Related Psychosis | 0.223 | 0.020 | 0.222 | 0.033 | 0.218 | 0.009 |
| ADHD | Substance Abuse | 0.250 | 0.015 | 0.246 | 0.020 | 0.235 | 0.005 |
| Benign Skin Neoplasm | ADHD | 0.215 | 0.010 | -0.116 | 0.015 | 0.089 | 0.005 |
| Non-Melanoma Skin Cancer | ADHD | 0.217 | 0.010 | -0.121 | 0.016 | 0.082 | 0.005 |
| Allergic Rhinitis | ADHD | 0.216 | 0.012 | 0.027 | 0.016 | 0.130 | 0.004 |
| Asthma | ADHD | 0.156 | 0.014 | 0.055 | 0.019 | 0.109 | 0.005 |
| Atopic Contact Dermatitis | ADHD | 0.221 | 0.019 | 0.019 | 0.015 | 0.091 | 0.005 |
| Chronic Sinusitis | ADHD | 0.174 | 0.012 | 0.041 | 0.020 | 0.120 | 0.005 |
| Cystitis Urethritis | ADHD | 0.123 | 0.019 | 0.064 | 0.020 | 0.085 | 0.007 |
| Esophagitis | ADHD | 0.161 | 0.023 | 0.029 | 0.028 | 0.093 | 0.008 |
| Eye Inflammation | ADHD | 0.156 | 0.016 | 0.009 | 0.016 | 0.074 | 0.005 |
| Fasciitis | ADHD | 0.104 | 0.019 | 0.022 | 0.025 | 0.060 | 0.007 |
| Gastritis Duodenitis | ADHD | 0.117 | 0.018 | 0.093 | 0.020 | 0.094 | 0.006 |
| IBS | ADHD | 0.203 | 0.022 | 0.018 | 0.023 | 0.098 | 0.007 |
| Lymphatic Disorder | ADHD | 0.100 | 0.033 | 0.038 | 0.031 | 0.056 | 0.009 |
| Osteoarthritis | ADHD | 0.200 | 0.018 | 0.042 | 0.020 | 0.103 | 0.007 |
| Psoriasis Related Disorders | ADHD | 0.101 | 0.025 | -0.010 | 0.031 | 0.042 | 0.008 |
| Type I Diabetes Mellitus | ADHD | -0.054 | 0.019 | -0.008 | 0.031 | -0.035 | 0.012 |
| Upper Respiratory Inflammation | ADHD | 0.221 | 0.018 | 0.014 | 0.019 | 0.104 | 0.006 |
| Cardiac Dysrhythmia | ADHD | 0.107 | 0.018 | 0.081 | 0.018 | 0.081 | 0.006 |
| General Hypertension | ADHD | 0.007 | 0.013 | 0.151 | 0.023 | 0.064 | 0.005 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Non-Rheumatic Heart Disease | ADHD | 0.054 | 0.016 | 0.032 | 0.020 | 0.040 | 0.007 |
| Adjustment Disorder | Anxiety Phobic Disorder | 0.592 | 0.012 | 0.211 | 0.010 | 0.385 | 0.003 |
| Adjustment Disorder | Bipolar Disorder | 0.394 | 0.012 | 0.240 | 0.015 | 0.327 | 0.006 |
| Adjustment Disorder | Depression | 0.671 | 0.010 | 0.242 | 0.009 | 0.450 | 0.003 |
| Adjustment Disorder | Migraine | 0.210 | 0.017 | 0.108 | 0.014 | 0.150 | 0.005 |
| Adjustment Disorder | Mood Disorder | 0.444 | 0.020 | 0.222 | 0.019 | 0.335 | 0.006 |
| Adjustment Disorder | Schizophrenia Related Psychosis | 0.314 | 0.018 | 0.199 | 0.019 | 0.259 | 0.007 |
| Adjustment Disorder | Substance Abuse | 0.353 | 0.014 | 0.111 | 0.012 | 0.222 | 0.004 |
| Benign Skin Neoplasm | Adjustment Disorder | 0.186 | 0.013 | 0.035 | 0.015 | 0.114 | 0.004 |
| Non-Melanoma Skin Cancer | Adjustment Disorder | 0.178 | 0.014 | 0.022 | 0.014 | 0.101 | 0.004 |
| Allergic Rhinitis | Adjustment Disorder | 0.141 | 0.013 | 0.065 | 0.011 | 0.100 | 0.004 |
| Asthma | Adjustment Disorder | 0.164 | 0.013 | 0.066 | 0.012 | 0.112 | 0.004 |
| Atopic Contact Dermatitis | Adjustment Disorder | 0.207 | 0.017 | 0.060 | 0.010 | 0.103 | 0.004 |
| Chronic Sinusitis | Adjustment Disorder | 0.162 | 0.012 | 0.060 | 0.014 | 0.112 | 0.004 |
| Cystitis Urethritis | Adjustment Disorder | 0.128 | 0.022 | 0.089 | 0.016 | 0.102 | 0.005 |
| Esophagitis | Adjustment Disorder | 0.182 | 0.023 | 0.039 | 0.019 | 0.099 | 0.006 |
| Eye Inflammation | Adjustment Disorder | 0.185 | 0.016 | 0.047 | 0.011 | 0.098 | 0.004 |
| Fasciitis | Adjustment Disorder | 0.179 | 0.020 | 0.032 | 0.015 | 0.092 | 0.005 |
| Gastritis Duodenitis | Adjustment Disorder | 0.145 | 0.021 | 0.093 | 0.014 | 0.110 | 0.004 |
| IBS | Adjustment Disorder | 0.286 | 0.023 | 0.062 | 0.015 | 0.143 | 0.006 |
| Lymphatic Disorder | Adjustment Disorder | 0.118 | 0.033 | 0.057 | 0.020 | 0.074 | 0.008 |
| Osteoarthritis | Adjustment Disorder | 0.192 | 0.020 | 0.087 | 0.012 | 0.120 | 0.004 |
| Psoriasis Related Disorders | Adjustment Disorder | 0.105 | 0.024 | 0.025 | 0.017 | 0.055 | 0.007 |
| Type I Diabetes Mellitus | Adjustment Disorder | 0.031 | 0.018 | 0.034 | 0.018 | 0.032 | 0.007 |
| Upper Respiratory Inflammation | Adjustment Disorder | 0.198 | 0.021 | 0.076 | 0.015 | 0.120 | 0.005 |
| Cardiac Dysrhythmia | Adjustment Disorder | 0.181 | 0.022 | 0.086 | 0.014 | 0.115 | 0.005 |
| General Hypertension | Adjustment Disorder | -0.068 | 0.012 | 0.142 | 0.012 | 0.041 | 0.004 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Non-Rheumatic Heart Disease | Adjustment Disorder | 0.051 | 0.022 | 0.073 | 0.017 | 0.062 | 0.006 |
| Anxiety Phobic Disorder | Bipolar Disorder | 0.598 | 0.016 | 0.393 | 0.018 | 0.487 | 0.004 |
| Anxiety Phobic Disorder | Depression | 0.738 | 0.011 | 0.425 | 0.010 | 0.570 | 0.003 |
| Anxiety Phobic Disorder | Migraine | 0.374 | 0.017 | 0.133 | 0.012 | 0.224 | 0.004 |
| Anxiety Phobic Disorder | Mood Disorder | 0.599 | 0.022 | 0.371 | 0.018 | 0.470 | 0.005 |
| Anxiety Phobic Disorder | Schizophrenia Related Psychosis | 0.441 | 0.026 | 0.429 | 0.024 | 0.432 | 0.006 |
| Anxiety Phobic Disorder | Substance Abuse | 0.513 | 0.015 | 0.263 | 0.010 | 0.363 | 0.003 |
| Benign Skin Neoplasm | Anxiety Phobic Disorder | 0.166 | 0.012 | 0.042 | 0.010 | 0.099 | 0.003 |
| Non-Melanoma Skin Cancer | Anxiety Phobic Disorder | 0.179 | 0.014 | 0.021 | 0.011 | 0.092 | 0.003 |
| Allergic Rhinitis | Anxiety Phobic Disorder | 0.258 | 0.012 | 0.109 | 0.008 | 0.171 | 0.003 |
| Asthma | Anxiety Phobic Disorder | 0.273 | 0.014 | 0.096 | 0.010 | 0.170 | 0.004 |
| Atopic Contact Dermatitis | Anxiety Phobic Disorder | 0.245 | 0.018 | 0.085 | 0.008 | 0.127 | 0.003 |
| Chronic Sinusitis | Anxiety Phobic Disorder | 0.252 | 0.013 | 0.111 | 0.010 | 0.174 | 0.003 |
| Cystitis Urethritis | Anxiety Phobic Disorder | 0.209 | 0.023 | 0.112 | 0.013 | 0.146 | 0.005 |
| Esophagitis | Anxiety Phobic Disorder | 0.287 | 0.026 | 0.122 | 0.016 | 0.182 | 0.005 |
| Eye Inflammation | Anxiety Phobic Disorder | 0.150 | 0.017 | 0.077 | 0.009 | 0.101 | 0.004 |
| Fasciitis | Anxiety Phobic Disorder | 0.168 | 0.020 | 0.076 | 0.013 | 0.109 | 0.005 |
| Gastritis Duodenitis | Anxiety Phobic Disorder | 0.289 | 0.019 | 0.200 | 0.010 | 0.228 | 0.003 |
| IBS | Anxiety Phobic Disorder | 0.386 | 0.024 | 0.209 | 0.013 | 0.265 | 0.005 |
| Lymphatic Disorder | Anxiety Phobic Disorder | 0.211 | 0.036 | 0.085 | 0.017 | 0.120 | 0.007 |
| Osteoarthritis | Anxiety Phobic Disorder | 0.218 | 0.018 | 0.140 | 0.009 | 0.162 | 0.003 |
| Psoriasis Related Disorders | Anxiety Phobic Disorder | 0.151 | 0.024 | 0.018 | 0.015 | 0.064 | 0.006 |
| Type I Diabetes Mellitus | Anxiety Phobic Disorder | 0.028 | 0.023 | 0.058 | 0.019 | 0.044 | 0.006 |
| Upper Respiratory Inflammation | Anxiety Phobic Disorder | 0.324 | 0.019 | 0.087 | 0.011 | 0.165 | 0.004 |
| Cardiac Dysrhythmia | Anxiety Phobic Disorder | 0.270 | 0.022 | 0.191 | 0.010 | 0.211 | 0.004 |
| General Hypertension | Anxiety Phobic Disorder | 0.103 | 0.013 | 0.216 | 0.010 | 0.168 | 0.003 |
| Non-Rheumatic Heart Disease | Anxiety Phobic Disorder | 0.110 | 0.021 | 0.163 | 0.013 | 0.143 | 0.004 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Bipolar Disorder | Depression | 0.711 | 0.012 | 0.499 | 0.018 | 0.622 | 0.004 |
| Bipolar Disorder | Migraine | 0.284 | 0.022 | 0.134 | 0.022 | 0.199 | 0.007 |
| Bipolar Disorder | Mood Disorder | 0.681 | 0.018 | 0.613 | 0.022 | 0.644 | 0.005 |
| Bipolar Disorder | Schizophrenia Related Psychosis | 0.698 | 0.019 | 0.625 | 0.026 | 0.664 | 0.005 |
| Bipolar Disorder | Substance Abuse | 0.490 | 0.021 | 0.444 | 0.024 | 0.456 | 0.005 |
| Benign Skin Neoplasm | Bipolar Disorder | 0.023 | 0.015 | 0.006 | 0.020 | 0.016 | 0.006 |
| Non-Melanoma Skin Cancer | Bipolar Disorder | 0.026 | 0.015 | 0.024 | 0.020 | 0.025 | 0.006 |
| Allergic Rhinitis | Bipolar Disorder | 0.127 | 0.016 | 0.056 | 0.018 | 0.091 | 0.006 |
| Asthma | Bipolar Disorder | 0.231 | 0.018 | 0.104 | 0.021 | 0.169 | 0.006 |
| Atopic Contact Dermatitis | Bipolar Disorder | 0.093 | 0.022 | 0.120 | 0.017 | 0.097 | 0.005 |
| Chronic Sinusitis | Bipolar Disorder | 0.170 | 0.016 | 0.083 | 0.022 | 0.131 | 0.005 |
| Cystitis Urethritis | Bipolar Disorder | 0.177 | 0.026 | 0.124 | 0.024 | 0.140 | 0.008 |
| Esophagitis | Bipolar Disorder | 0.212 | 0.029 | 0.097 | 0.029 | 0.146 | 0.008 |
| Eye Inflammation | Bipolar Disorder | 0.063 | 0.023 | 0.066 | 0.020 | 0.060 | 0.006 |
| Fasciitis | Bipolar Disorder | 0.071 | 0.023 | 0.097 | 0.022 | 0.080 | 0.008 |
| Gastritis Duodenitis | Bipolar Disorder | 0.258 | 0.025 | 0.162 | 0.021 | 0.191 | 0.006 |
| IBS | Bipolar Disorder | 0.282 | 0.028 | 0.130 | 0.024 | 0.183 | 0.008 |
| Lymphatic Disorder | Bipolar Disorder | 0.004 | 0.039 | 0.151 | 0.029 | 0.082 | 0.010 |
| Osteoarthritis | Bipolar Disorder | 0.236 | 0.023 | 0.088 | 0.019 | 0.139 | 0.006 |
| Psoriasis Related Disorders | Bipolar Disorder | 0.061 | 0.030 | 0.046 | 0.029 | 0.050 | 0.009 |
| Type I Diabetes Mellitus | Bipolar Disorder | 0.079 | 0.021 | 0.150 | 0.030 | 0.108 | 0.010 |
| Upper Respiratory Inflammation | Bipolar Disorder | 0.155 | 0.025 | 0.118 | 0.022 | 0.125 | 0.007 |
| Cardiac Dysrhythmia | Bipolar Disorder | 0.228 | 0.032 | 0.201 | 0.024 | 0.193 | 0.007 |
| General Hypertension | Bipolar Disorder | 0.126 | 0.017 | 0.199 | 0.021 | 0.156 | 0.006 |
| Non-Rheumatic Heart Disease | Bipolar Disorder | 0.094 | 0.023 | 0.108 | 0.024 | 0.096 | 0.008 |
| Depression | Migraine | 0.337 | 0.016 | 0.125 | 0.014 | 0.216 | 0.004 |
| Depression | Mood Disorder | 0.642 | 0.019 | 0.428 | 0.020 | 0.539 | 0.005 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Depression | Schizophrenia Related Psychosis | 0.511 | 0.017 | 0.495 | 0.020 | 0.503 | 0.005 |
| Depression | Substance Abuse | 0.435 | 0.014 | 0.331 | 0.013 | 0.376 | 0.004 |
| Benign Skin Neoplasm | Depression | 0.141 | 0.009 | -0.008 | 0.010 | 0.072 | 0.004 |
| Non-Melanoma Skin Cancer | Depression | 0.119 | 0.011 | 0.009 | 0.012 | 0.066 | 0.004 |
| Allergic Rhinitis | Depression | 0.137 | 0.012 | 0.108 | 0.011 | 0.121 | 0.003 |
| Asthma | Depression | 0.266 | 0.012 | 0.081 | 0.012 | 0.170 | 0.004 |
| Atopic Contact Dermatitis | Depression | 0.163 | 0.016 | 0.093 | 0.009 | 0.108 | 0.003 |
| Chronic Sinusitis | Depression | 0.205 | 0.011 | 0.101 | 0.012 | 0.155 | 0.004 |
| Cystitis Urethritis | Depression | 0.181 | 0.021 | 0.114 | 0.015 | 0.137 | 0.005 |
| Esophagitis | Depression | 0.199 | 0.022 | 0.129 | 0.019 | 0.156 | 0.006 |
| Eye Inflammation | Depression | 0.131 | 0.017 | 0.053 | 0.013 | 0.081 | 0.004 |
| Fasciitis | Depression | 0.143 | 0.018 | 0.062 | 0.014 | 0.095 | 0.004 |
| Gastritis Duodenitis | Depression | 0.219 | 0.019 | 0.191 | 0.013 | 0.194 | 0.004 |
| IBS | Depression | 0.342 | 0.023 | 0.142 | 0.015 | 0.214 | 0.005 |
| Lymphatic Disorder | Depression | 0.115 | 0.035 | 0.111 | 0.021 | 0.105 | 0.007 |
| Osteoarthritis | Depression | 0.232 | 0.019 | 0.143 | 0.012 | 0.168 | 0.004 |
| Psoriasis Related Disorders | Depression | 0.152 | 0.023 | -0.004 | 0.017 | 0.058 | 0.006 |
| Type I Diabetes Mellitus | Depression | 0.091 | 0.019 | 0.116 | 0.021 | 0.103 | 0.007 |
| Upper Respiratory Inflammation | Depression | 0.264 | 0.021 | 0.080 | 0.013 | 0.148 | 0.005 |
| Cardiac Dysrhythmia | Depression | 0.269 | 0.020 | 0.137 | 0.012 | 0.175 | 0.004 |
| General Hypertension | Depression | 0.048 | 0.012 | 0.220 | 0.013 | 0.135 | 0.003 |
| Non-Rheumatic Heart Disease | Depression | 0.041 | 0.017 | 0.179 | 0.014 | 0.115 | 0.004 |
| Migraine | Mood Disorder | 0.276 | 0.027 | 0.137 | 0.020 | 0.193 | 0.007 |
| Migraine | Schizophrenia Related Psychosis | 0.221 | 0.032 | 0.162 | 0.026 | 0.185 | 0.008 |
| Migraine | Substance Abuse | 0.218 | 0.020 | 0.074 | 0.012 | 0.128 | 0.005 |
| Benign Skin Neoplasm | Migraine | 0.132 | 0.015 | 0.080 | 0.012 | 0.101 | 0.004 |
| Non-Melanoma Skin Cancer | Migraine | 0.136 | 0.017 | 0.060 | 0.013 | 0.091 | 0.004 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Allergic Rhinitis | Migraine | 0.346 | 0.017 | 0.092 | 0.011 | 0.189 | 0.004 |
| Asthma | Migraine | 0.370 | 0.018 | 0.043 | 0.013 | 0.171 | 0.004 |
| Atopic Contact Dermatitis | Migraine | 0.282 | 0.022 | 0.063 | 0.008 | 0.119 | 0.004 |
| Chronic Sinusitis | Migraine | 0.334 | 0.017 | 0.157 | 0.012 | 0.229 | 0.004 |
| Cystitis Urethritis | Migraine | 0.209 | 0.029 | 0.096 | 0.015 | 0.132 | 0.005 |
| Esophagitis | Migraine | 0.365 | 0.031 | 0.057 | 0.018 | 0.163 | 0.006 |
| Eye Inflammation | Migraine | 0.234 | 0.021 | 0.074 | 0.011 | 0.124 | 0.004 |
| Fasciitis | Migraine | 0.253 | 0.028 | 0.034 | 0.015 | 0.109 | 0.005 |
| Gastritis Duodenitis | Migraine | 0.423 | 0.024 | 0.123 | 0.011 | 0.215 | 0.004 |
| IBS | Migraine | 0.478 | 0.030 | 0.084 | 0.015 | 0.206 | 0.005 |
| Lymphatic Disorder | Migraine | 0.372 | 0.055 | 0.011 | 0.021 | 0.102 | 0.008 |
| Osteoarthritis | Migraine | 0.414 | 0.028 | 0.060 | 0.013 | 0.163 | 0.004 |
| Psoriasis Related Disorders | Migraine | 0.156 | 0.030 | 0.017 | 0.016 | 0.061 | 0.007 |
| Type I Diabetes Mellitus | Migraine | 0.148 | 0.029 | -0.082 | 0.026 | 0.016 | 0.008 |
| Upper Respiratory Inflammation | Migraine | 0.372 | 0.025 | 0.130 | 0.013 | 0.205 | 0.005 |
| Cardiac Dysrhythmia | Migraine | 0.335 | 0.030 | 0.113 | 0.012 | 0.174 | 0.005 |
| General Hypertension | Migraine | 0.201 | 0.018 | 0.074 | 0.012 | 0.124 | 0.004 |
| Non-Rheumatic Heart Disease | Migraine | 0.273 | 0.028 | 0.082 | 0.015 | 0.146 | 0.005 |
| Mood Disorder | Schizophrenia Related Psychosis | 0.625 | 0.030 | 0.465 | 0.033 | 0.547 | 0.007 |
| Mood Disorder | Substance Abuse | 0.443 | 0.030 | 0.357 | 0.024 | 0.394 | 0.006 |
| Benign Skin Neoplasm | Mood Disorder | 0.047 | 0.019 | 0.049 | 0.020 | 0.048 | 0.006 |
| Non-Melanoma Skin Cancer | Mood Disorder | 0.029 | 0.023 | 0.051 | 0.024 | 0.040 | 0.007 |
| Allergic Rhinitis | Mood Disorder | 0.114 | 0.021 | 0.088 | 0.019 | 0.099 | 0.005 |
| Asthma | Mood Disorder | 0.237 | 0.023 | 0.052 | 0.024 | 0.138 | 0.007 |
| Atopic Contact Dermatitis | Mood Disorder | 0.150 | 0.032 | 0.070 | 0.017 | 0.091 | 0.006 |
| Chronic Sinusitis | Mood Disorder | 0.207 | 0.022 | 0.084 | 0.021 | 0.144 | 0.006 |
| Cystitis Urethritis | Mood Disorder | 0.185 | 0.034 | 0.100 | 0.024 | 0.130 | 0.010 |
| Esophagitis | Mood Disorder | 0.168 | 0.034 | 0.131 | 0.029 | 0.145 | 0.010 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Eye Inflammation | Mood Disorder | 0.100 | 0.027 | 0.056 | 0.019 | 0.070 | 0.007 |
| Fasciitis | Mood Disorder | 0.103 | 0.032 | 0.054 | 0.025 | 0.073 | 0.008 |
| Gastritis Duodenitis | Mood Disorder | 0.230 | 0.033 | 0.160 | 0.022 | 0.181 | 0.007 |
| IBS | Mood Disorder | 0.256 | 0.041 | 0.127 | 0.026 | 0.170 | 0.008 |
| Lymphatic Disorder | Mood Disorder | 0.152 | 0.047 | 0.097 | 0.029 | 0.111 | 0.012 |
| Osteoarthritis | Mood Disorder | 0.194 | 0.028 | 0.100 | 0.019 | 0.129 | 0.008 |
| Psoriasis Related Disorders | Mood Disorder | 0.170 | 0.047 | -0.011 | 0.034 | 0.057 | 0.011 |
| Type I Diabetes Mellitus | Mood Disorder | 0.115 | 0.029 | 0.078 | 0.029 | 0.097 | 0.013 |
| Upper Respiratory Inflammation | Mood Disorder | 0.227 | 0.031 | 0.105 | 0.021 | 0.147 | 0.008 |
| Cardiac Dysrhythmia | Mood Disorder | 0.184 | 0.030 | 0.236 | 0.019 | 0.209 | 0.007 |
| General Hypertension | Mood Disorder | 0.052 | 0.028 | 0.245 | 0.026 | 0.154 | 0.007 |
| Non-Rheumatic Heart Disease | Mood Disorder | 0.032 | 0.034 | 0.187 | 0.029 | 0.123 | 0.010 |
| Schizophrenia Related Psychosis | Substance Abuse | 0.332 | 0.019 | 0.503 | 0.018 | 0.422 | 0.006 |
| Benign Skin Neoplasm | Schizophrenia Related Psychosis | -0.008 | 0.020 | 0.013 | 0.022 | 0.002 | 0.007 |
| Non-Melanoma Skin Cancer | Schizophrenia Related Psychosis | -0.023 | 0.023 | 0.043 | 0.026 | 0.010 | 0.008 |
| Allergic Rhinitis | Schizophrenia Related Psychosis | 0.056 | 0.020 | 0.073 | 0.020 | 0.065 | 0.007 |
| Asthma | Schizophrenia Related Psychosis | 0.248 | 0.026 | 0.050 | 0.025 | 0.144 | 0.007 |
| Atopic Contact Dermatitis | Schizophrenia Related Psychosis | 0.105 | 0.033 | 0.091 | 0.019 | 0.089 | 0.006 |
| Chronic Sinusitis | Schizophrenia Related Psychosis | 0.043 | 0.022 | 0.169 | 0.025 | 0.104 | 0.007 |
| Cystitis Urethritis | Schizophrenia Related Psychosis | 0.192 | 0.042 | 0.052 | 0.033 | 0.106 | 0.010 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Esophagitis | Schizophrenia Related Psychosis | 0.196 | 0.045 | 0.116 | 0.036 | 0.148 | 0.010 |
| Eye Inflammation | Schizophrenia Related Psychosis | 0.073 | 0.029 | 0.064 | 0.021 | 0.065 | 0.008 |
| Fasciitis | Schizophrenia Related Psychosis | 0.031 | 0.036 | 0.085 | 0.030 | 0.061 | 0.010 |
| Gastritis Duodenitis | Schizophrenia Related Psychosis | 0.171 | 0.034 | 0.220 | 0.024 | 0.194 | 0.008 |
| IBS | Schizophrenia Related Psychosis | 0.125 | 0.045 | 0.142 | 0.027 | 0.140 | 0.011 |
| Lymphatic Disorder | Schizophrenia Related Psychosis | 0.231 | 0.056 | 0.083 | 0.037 | 0.128 | 0.013 |
| Osteoarthritis | Schizophrenia Related Psychosis | 0.159 | 0.033 | 0.136 | 0.021 | 0.138 | 0.008 |
| Psoriasis Related Disorders | Schizophrenia Related Psychosis | 0.019 | 0.041 | 0.082 | 0.031 | 0.054 | 0.011 |
| Type I Diabetes Mellitus | Schizophrenia Related Psychosis | 0.215 | 0.026 | 0.184 | 0.032 | 0.200 | 0.011 |
| Upper Respiratory Inflammation | Schizophrenia Related Psychosis | 0.059 | 0.033 | 0.164 | 0.024 | 0.118 | 0.008 |
| Cardiac Dysrhythmia | Schizophrenia Related Psychosis | 0.224 | 0.038 | 0.332 | 0.024 | 0.282 | 0.008 |
| General Hypertension | Schizophrenia Related Psychosis | 0.169 | 0.026 | 0.260 | 0.026 | 0.216 | 0.007 |
| Non-Rheumatic Heart Disease | Schizophrenia Related Psychosis | 0.153 | 0.036 | 0.208 | 0.027 | 0.181 | 0.009 |
| Benign Skin Neoplasm | Substance Abuse | -0.071 | 0.016 | -0.002 | 0.013 | -0.033 | 0.004 |
| Non-Melanoma Skin Cancer | Substance Abuse | -0.039 | 0.017 | 0.016 | 0.014 | -0.008 | 0.003 |
| Allergic Rhinitis | Substance Abuse | -0.007 | 0.015 | 0.040 | 0.010 | 0.021 | 0.003 |
| Asthma | Substance Abuse | 0.183 | 0.015 | 0.096 | 0.011 | 0.132 | 0.004 |
| Atopic Contact Dermatitis | Substance Abuse | 0.044 | 0.020 | 0.066 | 0.009 | 0.058 | 0.003 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Chronic Sinusitis | Substance Abuse | 0.108 | 0.015 | 0.079 | 0.012 | 0.091 | 0.004 |
| Cystitis Urethritis | Substance Abuse | 0.103 | 0.023 | 0.104 | 0.014 | 0.103 | 0.006 |
| Esophagitis | Substance Abuse | 0.159 | 0.025 | 0.123 | 0.016 | 0.136 | 0.006 |
| Eye Inflammation | Substance Abuse | -0.043 | 0.022 | 0.084 | 0.011 | 0.041 | 0.004 |
| Fasciitis | Substance Abuse | -0.021 | 0.024 | 0.012 | 0.015 | 0.000 | 0.005 |
| Gastritis Duodenitis | Substance Abuse | 0.175 | 0.022 | 0.188 | 0.012 | 0.182 | 0.004 |
| IBS | Substance Abuse | 0.146 | 0.027 | 0.076 | 0.015 | 0.098 | 0.006 |
| Lymphatic Disorder | Substance Abuse | 0.132 | 0.043 | 0.091 | 0.021 | 0.101 | 0.007 |
| Osteoarthritis | Substance Abuse | 0.228 | 0.023 | 0.077 | 0.011 | 0.123 | 0.004 |
| Psoriasis Related Disorders | Substance Abuse | 0.100 | 0.028 | 0.037 | 0.017 | 0.058 | 0.006 |
| Type I Diabetes Mellitus | Substance Abuse | 0.051 | 0.022 | 0.068 | 0.020 | 0.060 | 0.007 |
| Upper Respiratory Inflammation | Substance Abuse | 0.090 | 0.023 | 0.091 | 0.013 | 0.090 | 0.005 |
| Cardiac Dysrhythmia | Substance Abuse | 0.172 | 0.023 | 0.175 | 0.012 | 0.171 | 0.005 |
| General Hypertension | Substance Abuse | 0.076 | 0.016 | 0.211 | 0.012 | 0.154 | 0.004 |
| Non-Rheumatic Heart Disease | Substance Abuse | 0.016 | 0.025 | 0.165 | 0.015 | 0.111 | 0.005 |
| Benign Skin Neoplasm | Non-Melanoma Skin Cancer | 0.781 | 0.008 | 0.583 | 0.008 | 0.681 | 0.002 |
| Allergic Rhinitis | Benign Skin Neoplasm | 0.224 | 0.010 | 0.063 | 0.009 | 0.137 | 0.003 |
| Asthma | Benign Skin Neoplasm | 0.071 | 0.012 | 0.049 | 0.011 | 0.059 | 0.004 |
| Atopic Contact Dermatitis | Benign Skin Neoplasm | 0.365 | 0.015 | 0.125 | 0.008 | 0.193 | 0.003 |
| Chronic Sinusitis | Benign Skin Neoplasm | 0.196 | 0.012 | 0.045 | 0.012 | 0.122 | 0.003 |
| Cystitis Urethritis | Benign Skin Neoplasm | 0.122 | 0.020 | 0.059 | 0.015 | 0.082 | 0.005 |
| Esophagitis | Benign Skin Neoplasm | 0.224 | 0.020 | 0.024 | 0.016 | 0.108 | 0.005 |
| Eye Inflammation | Benign Skin Neoplasm | 0.302 | 0.013 | 0.059 | 0.009 | 0.150 | 0.003 |
| Fasciitis | Benign Skin Neoplasm | 0.205 | 0.016 | 0.068 | 0.012 | 0.123 | 0.004 |
| Gastritis Duodenitis | Benign Skin Neoplasm | 0.103 | 0.018 | 0.076 | 0.011 | 0.084 | 0.003 |
| IBS | Benign Skin Neoplasm | 0.334 | 0.022 | 0.008 | 0.015 | 0.131 | 0.005 |
| Lymphatic Disorder | Benign Skin Neoplasm | 0.144 | 0.032 | 0.056 | 0.018 | 0.082 | 0.006 |
| Osteoarthritis | Benign Skin Neoplasm | 0.164 | 0.017 | 0.084 | 0.010 | 0.109 | 0.003 |
| Psoriasis Related Disorders | Benign Skin Neoplasm | 0.232 | 0.023 | 0.056 | 0.015 | 0.122 | 0.005 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Type I Diabetes Mellitus | Benign Skin Neoplasm | -0.154 | 0.019 | 0.076 | 0.020 | -0.041 | 0.006 |
| Upper Respiratory Inflammation | Benign Skin Neoplasm | 0.280 | 0.018 | 0.043 | 0.012 | 0.131 | 0.004 |
| Cardiac Dysrhythmia | Benign Skin Neoplasm | 0.163 | 0.019 | 0.053 | 0.011 | 0.088 | 0.003 |
| General Hypertension | Benign Skin Neoplasm | -0.132 | 0.011 | 0.167 | 0.011 | 0.023 | 0.003 |
| Non-Rheumatic Heart Disease | Benign Skin Neoplasm | 0.158 | 0.016 | 0.052 | 0.011 | 0.095 | 0.004 |
| Allergic Rhinitis | Non-Melanoma Skin Cancer | 0.220 | 0.011 | 0.039 | 0.010 | 0.121 | 0.003 |
| Asthma | Non-Melanoma Skin Cancer | 0.046 | 0.014 | 0.051 | 0.012 | 0.049 | 0.004 |
| Atopic Contact Dermatitis | Non-Melanoma Skin Cancer | 0.287 | 0.015 | 0.179 | 0.008 | 0.202 | 0.003 |
| Chronic Sinusitis | Non-Melanoma Skin Cancer | 0.164 | 0.012 | 0.052 | 0.012 | 0.107 | 0.004 |
| Cystitis Urethritis | Non-Melanoma Skin Cancer | 0.186 | 0.023 | 0.028 | 0.016 | 0.087 | 0.005 |
| Esophagitis | Non-Melanoma Skin Cancer | 0.168 | 0.025 | 0.045 | 0.019 | 0.095 | 0.005 |
| Eye Inflammation | Non-Melanoma Skin Cancer | 0.221 | 0.016 | 0.078 | 0.010 | 0.129 | 0.003 |
| Fasciitis | Non-Melanoma Skin Cancer | 0.189 | 0.022 | 0.054 | 0.015 | 0.108 | 0.003 |
| Gastritis Duodenitis | Non-Melanoma Skin Cancer | 0.109 | 0.018 | 0.065 | 0.012 | 0.079 | 0.004 |
| IBS | Non-Melanoma Skin Cancer | 0.245 | 0.025 | 0.035 | 0.015 | 0.109 | 0.005 |
| Lymphatic Disorder | Non-Melanoma Skin Cancer | 0.210 | 0.040 | 0.028 | 0.019 | 0.083 | 0.006 |
| Osteoarthritis | Non-Melanoma Skin Cancer | 0.159 | 0.019 | 0.080 | 0.011 | 0.105 | 0.003 |
| Psoriasis Related Disorders | Non-Melanoma Skin Cancer | 0.173 | 0.025 | 0.110 | 0.017 | 0.131 | 0.006 |
| Type I Diabetes Mellitus | Non-Melanoma Skin Cancer | -0.122 | 0.020 | 0.043 | 0.022 | -0.041 | 0.006 |
| Upper Respiratory Inflammation | Non-Melanoma Skin Cancer | 0.240 | 0.020 | 0.045 | 0.013 | 0.115 | 0.004 |
| Cardiac Dysrhythmia | Non-Melanoma Skin Cancer | 0.108 | 0.021 | 0.080 | 0.011 | 0.086 | 0.004 |
| General Hypertension | Non-Melanoma Skin Cancer | -0.113 | 0.014 | 0.145 | 0.012 | 0.025 | 0.003 |
| Non-Rheumatic Heart Disease | Non-Melanoma Skin Cancer | 0.145 | 0.020 | 0.051 | 0.015 | 0.088 | 0.004 |
| Allergic Rhinitis | Asthma | 0.493 | 0.011 | 0.345 | 0.008 | 0.407 | 0.003 |
| Allergic Rhinitis | Atopic Contact Dermatitis | 0.434 | 0.016 | 0.128 | 0.006 | 0.210 | 0.003 |
| Allergic Rhinitis | Chronic Sinusitis | 0.479 | 0.010 | 0.368 | 0.009 | 0.418 | 0.003 |
| Allergic Rhinitis | Cystitis Urethritis | 0.328 | 0.021 | 0.027 | 0.012 | 0.132 | 0.005 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Allergic Rhinitis | Esophagitis | 0.346 | 0.024 | 0.061 | 0.015 | 0.169 | 0.005 |
| Allergic Rhinitis | Eye Inflammation | 0.376 | 0.015 | 0.110 | 0.008 | 0.200 | 0.003 |
| Allergic Rhinitis | Fasciitis | 0.315 | 0.018 | 0.064 | 0.010 | 0.157 | 0.004 |
| Allergic Rhinitis | Gastritis Duodenitis | 0.440 | 0.017 | 0.064 | 0.009 | 0.191 | 0.003 |
| Allergic Rhinitis | IBS | 0.451 | 0.023 | 0.066 | 0.012 | 0.194 | 0.005 |
| Allergic Rhinitis | Lymphatic Disorder | 0.289 | 0.035 | 0.035 | 0.015 | 0.108 | 0.006 |
| Allergic Rhinitis | Osteoarthritis | 0.359 | 0.018 | 0.075 | 0.008 | 0.165 | 0.003 |
| Allergic Rhinitis | Psoriasis Related Disorders | 0.133 | 0.025 | 0.042 | 0.014 | 0.073 | 0.005 |
| Allergic Rhinitis | Type I Diabetes Mellitus | 0.089 | 0.020 | -0.059 | 0.018 | 0.009 | 0.006 |
| Allergic Rhinitis | Upper Respiratory Inflammation | 0.481 | 0.019 | 0.280 | 0.010 | 0.344 | 0.003 |
| Cardiac Dysrhythmia | Allergic Rhinitis | 0.156 | 0.019 | 0.068 | 0.009 | 0.094 | 0.004 |
| General Hypertension | Allergic Rhinitis | 0.210 | 0.011 | 0.080 | 0.009 | 0.136 | 0.003 |
| Non-Rheumatic Heart Disease | Allergic Rhinitis | 0.259 | 0.018 | 0.035 | 0.011 | 0.118 | 0.004 |
| Asthma | Atopic Contact Dermatitis | 0.326 | 0.017 | 0.085 | 0.008 | 0.151 | 0.003 |
| Asthma | Chronic Sinusitis | 0.358 | 0.012 | 0.212 | 0.011 | 0.279 | 0.003 |
| Asthma | Cystitis Urethritis | 0.228 | 0.024 | 0.020 | 0.015 | 0.095 | 0.005 |
| Asthma | Esophagitis | 0.328 | 0.026 | 0.084 | 0.017 | 0.179 | 0.005 |
| Asthma | Eye Inflammation | 0.214 | 0.017 | 0.086 | 0.009 | 0.129 | 0.004 |
| Asthma | Fasciitis | 0.233 | 0.021 | 0.073 | 0.015 | 0.133 | 0.005 |
| Asthma | Gastritis Duodenitis | 0.383 | 0.020 | 0.091 | 0.011 | 0.190 | 0.004 |
| Asthma | IBS | 0.392 | 0.027 | 0.029 | 0.015 | 0.152 | 0.005 |
| Asthma | Lymphatic Disorder | 0.346 | 0.044 | 0.041 | 0.017 | 0.128 | 0.007 |
| Asthma | Osteoarthritis | 0.335 | 0.020 | 0.122 | 0.011 | 0.188 | 0.004 |
| Asthma | Psoriasis Related Disorders | 0.136 | 0.029 | 0.019 | 0.016 | 0.060 | 0.006 |
| Asthma | Type I Diabetes Mellitus | 0.215 | 0.022 | 0.008 | 0.021 | 0.104 | 0.006 |
| Asthma | Upper Respiratory Inflammation | 0.370 | 0.021 | 0.205 | 0.011 | 0.258 | 0.004 |
| Cardiac Dysrhythmia | Asthma | 0.311 | 0.023 | 0.110 | 0.012 | 0.170 | 0.004 |
| General Hypertension | Asthma | 0.263 | 0.014 | 0.101 | 0.011 | 0.172 | 0.003 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Non-Rheumatic Heart Disease | Asthma | 0.239 | 0.022 | 0.107 | 0.013 | 0.155 | 0.005 |
| Atopic Contact Dermatitis | Chronic Sinusitis | 0.337 | 0.018 | 0.091 | 0.009 | 0.161 | 0.003 |
| Atopic Contact Dermatitis | Cystitis Urethritis | 0.256 | 0.029 | 0.077 | 0.010 | 0.118 | 0.004 |
| Atopic Contact Dermatitis | Esophagitis | 0.268 | 0.034 | 0.064 | 0.013 | 0.116 | 0.005 |
| Atopic Contact Dermatitis | Eye Inflammation | 0.398 | 0.024 | 0.102 | 0.007 | 0.169 | 0.003 |
| Atopic Contact Dermatitis | Fasciitis | 0.380 | 0.027 | 0.059 | 0.009 | 0.139 | 0.004 |
| Atopic Contact Dermatitis | Gastritis Duodenitis | 0.364 | 0.025 | 0.067 | 0.008 | 0.134 | 0.003 |
| Atopic Contact Dermatitis | IBS | 0.398 | 0.035 | 0.070 | 0.011 | 0.143 | 0.004 |
| Atopic Contact Dermatitis | Lymphatic Disorder | 0.324 | 0.047 | 0.072 | 0.013 | 0.122 | 0.006 |
| Atopic Contact Dermatitis | Osteoarthritis | 0.350 | 0.024 | 0.096 | 0.008 | 0.151 | 0.003 |
| Atopic Contact Dermatitis | Psoriasis Related Disorders | 0.355 | 0.029 | 0.237 | 0.011 | 0.263 | 0.005 |
| Atopic Contact Dermatitis | Type I Diabetes Mellitus | 0.066 | 0.028 | 0.022 | 0.015 | 0.034 | 0.006 |
| Atopic Contact Dermatitis | Upper Respiratory Inflammation | 0.372 | 0.028 | 0.087 | 0.009 | 0.151 | 0.004 |
| Cardiac Dysrhythmia | Atopic Contact Dermatitis | 0.303 | 0.030 | 0.052 | 0.009 | 0.104 | 0.004 |
| General Hypertension | Atopic Contact Dermatitis | 0.143 | 0.016 | 0.073 | 0.008 | 0.090 | 0.003 |
| Non-Rheumatic Heart Disease | Atopic Contact Dermatitis | 0.159 | 0.028 | 0.068 | 0.010 | 0.090 | 0.004 |
| Chronic Sinusitis | Cystitis Urethritis | 0.215 | 0.021 | 0.056 | 0.014 | 0.115 | 0.005 |
| Chronic Sinusitis | Esophagitis | 0.330 | 0.022 | 0.074 | 0.017 | 0.180 | 0.005 |
| Chronic Sinusitis | Eye Inflammation | 0.261 | 0.016 | 0.106 | 0.010 | 0.161 | 0.004 |
| Chronic Sinusitis | Fasciitis | 0.280 | 0.020 | 0.055 | 0.014 | 0.145 | 0.005 |
| Chronic Sinusitis | Gastritis Duodenitis | 0.378 | 0.018 | 0.106 | 0.012 | 0.203 | 0.004 |
| Chronic Sinusitis | IBS | 0.391 | 0.022 | 0.077 | 0.013 | 0.188 | 0.005 |
| Chronic Sinusitis | Lymphatic Disorder | 0.327 | 0.037 | 0.069 | 0.019 | 0.148 | 0.007 |
| Chronic Sinusitis | Osteoarthritis | 0.389 | 0.020 | 0.081 | 0.012 | 0.185 | 0.004 |
| Chronic Sinusitis | Psoriasis Related Disorders | 0.145 | 0.024 | 0.016 | 0.017 | 0.065 | 0.006 |
| Chronic Sinusitis | Type I Diabetes Mellitus | 0.114 | 0.020 | 0.024 | 0.021 | 0.069 | 0.006 |
| Chronic Sinusitis | Upper Respiratory Inflammation | 0.563 | 0.019 | 0.409 | 0.011 | 0.454 | 0.004 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Cardiac Dysrhythmia | Chronic Sinusitis | 0.293 | 0.022 | 0.088 | 0.012 | 0.153 | 0.004 |
| General Hypertension | Chronic Sinusitis | 0.201 | 0.013 | 0.100 | 0.012 | 0.147 | 0.003 |
| Non-Rheumatic Heart Disease | Chronic Sinusitis | 0.217 | 0.019 | 0.080 | 0.015 | 0.134 | 0.004 |
| Cystitis Urethritis | Esophagitis | 0.269 | 0.038 | 0.071 | 0.021 | 0.135 | 0.008 |
| Cystitis Urethritis | Eye Inflammation | 0.140 | 0.030 | 0.071 | 0.012 | 0.091 | 0.005 |
| Cystitis Urethritis | Fasciitis | 0.151 | 0.036 | 0.050 | 0.018 | 0.081 | 0.005 |
| Cystitis Urethritis | Gastritis Duodenitis | 0.311 | 0.035 | 0.114 | 0.016 | 0.169 | 0.005 |
| Cystitis Urethritis | IBS | 0.344 | 0.042 | 0.102 | 0.018 | 0.171 | 0.007 |
| Cystitis Urethritis | Lymphatic Disorder | 0.214 | 0.065 | 0.044 | 0.023 | 0.085 | 0.009 |
| Cystitis Urethritis | Osteoarthritis | 0.158 | 0.037 | 0.095 | 0.015 | 0.112 | 0.005 |
| Cystitis Urethritis | Psoriasis Related Disorders | 0.052 | 0.038 | 0.037 | 0.020 | 0.041 | 0.008 |
| Cystitis Urethritis | Type I Diabetes Mellitus | 0.066 | 0.038 | 0.078 | 0.027 | 0.072 | 0.009 |
| Cystitis Urethritis | Upper Respiratory Inflammation | 0.273 | 0.033 | 0.058 | 0.015 | 0.119 | 0.006 |
| Cardiac Dysrhythmia | Cystitis Urethritis | 0.184 | 0.037 | 0.086 | 0.014 | 0.111 | 0.006 |
| General Hypertension | Cystitis Urethritis | 0.156 | 0.021 | 0.067 | 0.014 | 0.099 | 0.005 |
| Non-Rheumatic Heart Disease | Cystitis Urethritis | 0.213 | 0.037 | 0.073 | 0.019 | 0.117 | 0.007 |
| Esophagitis | Eye Inflammation | 0.243 | 0.028 | 0.044 | 0.014 | 0.106 | 0.005 |
| Esophagitis | Fasciitis | 0.256 | 0.039 | 0.056 | 0.021 | 0.123 | 0.006 |
| Esophagitis | Gastritis Duodenitis | 0.679 | 0.031 | 0.622 | 0.014 | 0.637 | 0.003 |
| Esophagitis | IBS | 0.404 | 0.045 | 0.242 | 0.022 | 0.290 | 0.006 |
| Esophagitis | Lymphatic Disorder | 0.275 | 0.054 | 0.051 | 0.024 | 0.112 | 0.010 |
| Esophagitis | Osteoarthritis | 0.398 | 0.033 | 0.074 | 0.014 | 0.167 | 0.005 |
| Esophagitis | Psoriasis Related Disorders | 0.090 | 0.042 | 0.053 | 0.022 | 0.065 | 0.009 |
| Esophagitis | Type I Diabetes Mellitus | 0.162 | 0.042 | 0.007 | 0.033 | 0.071 | 0.009 |
| Esophagitis | Upper Respiratory Inflammation | 0.347 | 0.038 | 0.101 | 0.018 | 0.176 | 0.006 |
| Cardiac Dysrhythmia | Esophagitis | 0.423 | 0.039 | 0.064 | 0.016 | 0.165 | 0.006 |
| General Hypertension | Esophagitis | 0.289 | 0.026 | 0.060 | 0.019 | 0.151 | 0.005 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Non-Rheumatic Heart Disease | Esophagitis | 0.352 | 0.039 | 0.087 | 0.021 | 0.175 | 0.006 |
| Eye Inflammation | Fasciitis | 0.271 | 0.025 | 0.039 | 0.012 | 0.109 | 0.005 |
| Eye Inflammation | Gastritis Duodenitis | 0.276 | 0.024 | 0.068 | 0.010 | 0.125 | 0.004 |
| Eye Inflammation | IBS | 0.286 | 0.031 | 0.067 | 0.013 | 0.126 | 0.005 |
| Eye Inflammation | Lymphatic Disorder | 0.152 | 0.039 | 0.066 | 0.014 | 0.086 | 0.007 |
| Eye Inflammation | Osteoarthritis | 0.192 | 0.026 | 0.088 | 0.010 | 0.115 | 0.004 |
| Eye Inflammation | Psoriasis Related Disorders | 0.224 | 0.033 | 0.040 | 0.015 | 0.091 | 0.006 |
| Eye Inflammation | Type I Diabetes Mellitus | 0.007 | 0.031 | 0.094 | 0.020 | 0.060 | 0.007 |
| Eye Inflammation | Upper Respiratory Inflammation | 0.313 | 0.026 | 0.091 | 0.011 | 0.152 | 0.004 |
| Cardiac Dysrhythmia | Eye Inflammation | 0.184 | 0.025 | 0.061 | 0.009 | 0.091 | 0.004 |
| General Hypertension | Eye Inflammation | 0.074 | 0.016 | 0.069 | 0.010 | 0.070 | 0.004 |
| Non-Rheumatic Heart Disease | Eye Inflammation | 0.218 | 0.028 | 0.050 | 0.013 | 0.100 | 0.005 |
| Fasciitis | Gastritis Duodenitis | 0.258 | 0.029 | 0.079 | 0.013 | 0.132 | 0.004 |
| Fasciitis | IBS | 0.301 | 0.034 | 0.044 | 0.016 | 0.121 | 0.006 |
| Fasciitis | Lymphatic Disorder | 0.177 | 0.046 | 0.073 | 0.020 | 0.099 | 0.008 |
| Fasciitis | Osteoarthritis | 0.489 | 0.031 | 0.162 | 0.013 | 0.254 | 0.004 |
| Fasciitis | Psoriasis Related Disorders | 0.209 | 0.036 | 0.010 | 0.019 | 0.072 | 0.007 |
| Fasciitis | Type I Diabetes Mellitus | 0.098 | 0.033 | 0.085 | 0.025 | 0.089 | 0.007 |
| Fasciitis | Upper Respiratory Inflammation | 0.327 | 0.031 | 0.044 | 0.013 | 0.129 | 0.005 |
| Cardiac Dysrhythmia | Fasciitis | 0.157 | 0.035 | 0.061 | 0.014 | 0.086 | 0.005 |
| General Hypertension | Fasciitis | 0.113 | 0.021 | 0.122 | 0.013 | 0.118 | 0.004 |
| Non-Rheumatic Heart Disease | Fasciitis | 0.087 | 0.033 | 0.079 | 0.017 | 0.081 | 0.005 |
| Gastritis Duodenitis | IBS | 0.508 | 0.034 | 0.322 | 0.013 | 0.371 | 0.005 |
| Gastritis Duodenitis | Lymphatic Disorder | 0.376 | 0.043 | 0.059 | 0.014 | 0.134 | 0.007 |
| Gastritis Duodenitis | Osteoarthritis | 0.457 | 0.030 | 0.116 | 0.011 | 0.204 | 0.004 |
| Gastritis Duodenitis | Psoriasis Related Disorders | 0.155 | 0.041 | 0.037 | 0.017 | 0.070 | 0.007 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Gastritis Duodenitis | Type I Diabetes Mellitus | 0.200 | 0.034 | 0.114 | 0.022 | 0.142 | 0.006 |
| Gastritis Duodenitis | Upper Respiratory Inflammation | 0.450 | 0.029 | 0.100 | 0.012 | 0.195 | 0.005 |
| Cardiac Dysrhythmia | Gastritis Duodenitis | 0.450 | 0.030 | 0.131 | 0.011 | 0.210 | 0.004 |
| General Hypertension | Gastritis Duodenitis | 0.427 | 0.020 | 0.094 | 0.011 | 0.207 | 0.004 |
| Non-Rheumatic Heart Disease | Gastritis Duodenitis | 0.456 | 0.027 | 0.114 | 0.012 | 0.216 | 0.004 |
| IBS | Lymphatic Disorder | 0.329 | 0.053 | 0.040 | 0.019 | 0.108 | 0.009 |
| IBS | Osteoarthritis | 0.402 | 0.038 | 0.075 | 0.014 | 0.158 | 0.005 |
| IBS | Psoriasis Related Disorders | 0.228 | 0.052 | 0.023 | 0.021 | 0.080 | 0.008 |
| IBS | Type I Diabetes Mellitus | 0.062 | 0.041 | -0.004 | 0.025 | 0.007 | 0.011 |
| IBS | Upper Respiratory Inflammation | 0.389 | 0.037 | 0.078 | 0.015 | 0.162 | 0.006 |
| Cardiac Dysrhythmia | IBS | 0.298 | 0.038 | 0.093 | 0.014 | 0.142 | 0.006 |
| General Hypertension | IBS | 0.188 | 0.027 | 0.040 | 0.016 | 0.089 | 0.005 |
| Non-Rheumatic Heart Disease | IBS | 0.311 | 0.033 | 0.064 | 0.016 | 0.137 | 0.006 |
| Lymphatic Disorder | Osteoarthritis | 0.332 | 0.056 | 0.111 | 0.017 | 0.160 | 0.007 |
| Lymphatic Disorder | Psoriasis Related Disorders | 0.129 | 0.040 | 0.045 | 0.020 | 0.065 | 0.011 |
| Lymphatic Disorder | Type I Diabetes Mellitus | 0.188 | 0.052 | 0.133 | 0.028 | 0.143 | 0.011 |
| Lymphatic Disorder | Upper Respiratory Inflammation | 0.339 | 0.056 | 0.078 | 0.019 | 0.140 | 0.008 |
| Cardiac Dysrhythmia | Lymphatic Disorder | 0.318 | 0.052 | 0.096 | 0.015 | 0.142 | 0.007 |
| General Hypertension | Lymphatic Disorder | 0.226 | 0.038 | 0.092 | 0.018 | 0.128 | 0.007 |
| Non-Rheumatic Heart Disease | Lymphatic Disorder | 0.338 | 0.056 | 0.068 | 0.021 | 0.137 | 0.008 |
| Osteoarthritis | Psoriasis Related Disorders | 0.126 | 0.036 | 0.121 | 0.014 | 0.122 | 0.006 |
| Osteoarthritis | Type I Diabetes Mellitus | 0.231 | 0.037 | 0.097 | 0.022 | 0.139 | 0.006 |
| Osteoarthritis | Upper Respiratory Inflammation | 0.412 | 0.028 | 0.084 | 0.011 | 0.168 | 0.004 |
| Cardiac Dysrhythmia | Osteoarthritis | 0.362 | 0.032 | 0.124 | 0.010 | 0.180 | 0.004 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| General Hypertension | Osteoarthritis | 0.404 | 0.022 | 0.152 | 0.010 | 0.231 | 0.003 |
| Non-Rheumatic Heart Disease | Osteoarthritis | 0.278 | 0.037 | 0.106 | 0.015 | 0.153 | 0.004 |
| Psoriasis Related Disorders | Type I Diabetes Mellitus | 0.161 | 0.037 | -0.051 | 0.027 | 0.031 | 0.010 |
| Psoriasis Related Disorders | Upper Respiratory Inflammation | 0.161 | 0.035 | 0.037 | 0.015 | 0.072 | 0.007 |
| Cardiac Dysrhythmia | Psoriasis Related Disorders | 0.061 | 0.041 | 0.062 | 0.017 | 0.061 | 0.006 |
| General Hypertension | Psoriasis Related Disorders | 0.075 | 0.027 | 0.078 | 0.017 | 0.076 | 0.006 |
| Non-Rheumatic Heart Disease | Psoriasis Related Disorders | 0.103 | 0.039 | 0.060 | 0.019 | 0.073 | 0.007 |
| Type I Diabetes Mellitus | Upper Respiratory Inflammation | 0.113 | 0.032 | 0.014 | 0.023 | 0.051 | 0.008 |
| Cardiac Dysrhythmia | Type I Diabetes Mellitus | 0.233 | 0.032 | 0.167 | 0.019 | 0.182 | 0.007 |
| General Hypertension | Type I Diabetes Mellitus | 0.467 | 0.021 | 0.361 | 0.020 | 0.410 | 0.005 |
| Non-Rheumatic Heart Disease | Type I Diabetes Mellitus | 0.192 | 0.029 | 0.186 | 0.022 | 0.185 | 0.006 |
| Cardiac Dysrhythmia | Upper Respiratory Inflammation | 0.319 | 0.031 | 0.094 | 0.012 | 0.149 | 0.005 |
| General Hypertension | Upper Respiratory Inflammation | 0.156 | 0.020 | 0.109 | 0.011 | 0.124 | 0.004 |
| Non-Rheumatic Heart Disease | Upper Respiratory Inflammation | 0.231 | 0.033 | 0.106 | 0.014 | 0.143 | 0.005 |
| Cardiac Dysrhythmia | General Hypertension | 0.378 | 0.022 | 0.251 | 0.010 | 0.284 | 0.003 |
| Cardiac Dysrhythmia | Non-Rheumatic Heart Disease | 0.558 | 0.029 | 0.457 | 0.012 | 0.482 | 0.004 |
| General Hypertension | Non-Rheumatic Heart Disease | 0.400 | 0.021 | 0.277 | 0.013 | 0.321 | 0.004 |

**Table A.5: Genetic correlation estimates compared to published GWAS studies. Bold** font indicates the two estimates' 95% confidence intervals do not overlap with each other

| **Diease1** | **Diesase2** | **Family $r_g$ (SD)** | **GWAS $r_g$ (SE)** | **GWAS Study** |
|---|---|---|---|---|
| ADHD | Bipolar Disorder | 0.37(0.01) | 0.25(0.06) | [7] |
| ADHD | Depression | 0.42(0.01) | 0.48(0.1) | [7] |
| ADHD | Migraine | 0.21(0.02) | 0.26(0.05) | [7] |
| ADHD | Schizophrenia / Related Psychosis | 0.22(0.02) | 0.22(0.05) | [7] |
| ADHD | Type I Diabetes Mellitus | -0.05(0.02) | -0.04(0.13) | [35] |
| ADHD | Non-Rheumatic Heart Disease | 0.05(0.02) | 0.22(0.08) | [7] |
| Bipolar Disorder | Depression | 0.71(0.01) | 0.47(0.05) | [52] |
| Bipolar Disorder | Migraine | 0.28(0.02) | -0.02(0.03) | [7] |
| Bipolar Disorder | Schizophrenia / Related Psychosis | 0.70(0.02) | 0.68(0.02) | [7] |
| Bipolar Disorder | Type I Diabetes Mellitus | 0.08(0.02) | 0.01(0.09) | [35] |

Continued on next page

**Table A.5 – continued from previous page**

| Diease1 | Diesase2 | Family $r_g$ (SD) | GWAS $r_g$ (SE) | GWAS Study |
|---|---|---|---|---|
| Bipolar Disorder | Non-Rheumatic Heart Disease | 0.09(0.02) | -0.11(0.07) | [35] |
| Depression | Migraine | 0.34(0.02) | 0.21(0.05) | [7] |
| Depression | Schizophrenia / Related Psychosis | 0.51(0.02) | 0.51(0.07) | [35] |
| Depression | Type I Diabetes Mellitus | 0.09(0.02) | -0.05(0.11) | [35] |
| Depression | Non-Rheumatic Heart Disease | 0.04(0.02) | 0.03(0.12) | [35] |
| Migraine | Schizophrenia / Related Psychosis | 0.22(0.03) | -0.09(0.03) | [7] |
| Migraine | Non-Rheumatic Heart Disease | 0.27(0.03) | 0.05(0.06) | [7] |
| Schizophrenia / Related Psychosis | Type I Diabetes Mellitus | 0.21(0.03) | -0.04(0.04) | [35] |
| Schizophrenia / Related Psychosis | Non-Rheumatic Heart Disease | 0.15(0.04) | -0.0(0.04) | [35] |

Continued on next page

**Table A.5 – continued from previous page**

| Diease1 | Diesase2 | Family $r_g$ (SD) | GWAS $r_g$ (SE) | GWAS Study |
|---|---|---|---|---|
| Non-Rheumatic Heart Disease | Type I Diabetes Mellitus | 0.19(0.03) | 0.15(0.09) | [35] |
| Allergic Rhinitis | Asthma | 0.49(0.01) | 0.85(0.1) | [175] |
| Allergic Rhinitis | Osteoarthritis | 0.36(0.02) | 0.39(0.12) | [175] |
| Allergic Rhinitis | General Hypertension | 0.21(0.01) | 0.23(0.08) | [175] |
| Allergic Rhinitis | Non-Rheumatic Heart Disease | 0.26(0.02) | -0.04(0.13) | [175] |
| Asthma | Osteoarthritis | 0.33(0.02) | 0.35(0.09) | [175] |
| Asthma | General Hypertension | 0.26(0.01) | 0.27(0.06) | [175] |
| Asthma | Non-Rheumatic Heart Disease | 0.24(0.02) | 0.29(0.07) | [175] |
| Osteoarthritis | General Hypertension | 0.40(0.02) | 0.22(0.06) | [175] |
| Osteoarthritis | Non-Rheumatic Heart Disease | 0.28(0.04) | 0.31(0.11) | [175] |
| General Hypertension | Non-Rheumatic Heart Disease | 0.40(0.02) | 0.53(0.07) | [175] |

**Table A.6: Heritability estimates compared to published GWAS studies. Bold** font indicates the two estimates' 95% confidence intervals do not overlap with each other

| **Disease** | **Family $h^2$ Adjusted** | **Family $h^2$ SD** | **GWAS $h^2$** | **GWAS Reference** |
|---|---|---|---|---|
| Epilepsy Related Disorders | 0.541 | 0.019 | 0.32 | [266] |
| Cardiac Dysrhythmia | 0.24 | 0.011 | 0.058 | [259] |
| Cerebrovascular Disease | 0.217 | 0.021 | 0.09 | [199] |
| General Hypertension | 0.462 | 0.009 | 0.32 | [199] |
| Goiter | 0.408 | 0.017 | 0.056 | [227] |
| Type II Diabetes Mellitus | 0.561 | 0.01 | 0.35 | [199] |
| Biliary Tract Disease | 0.219 | 0.013 | 0.063 | [173] |
| Asthma | 0.457 | 0.008 | 0.106 | [193] |
| Atopic Contact Dermatitis | 0.202 | 0.006 | 0.149 | [281] |
| Crohn's Disease | 0.574 | 0.028 | 0.284 | [178] |
| Inflammatory Spondylopathies | 0.646 | 0.016 | 0.237 | [214] |
| Lupus Erythematosus | 0.66 | 0.027 | 0.414 | [325] |
| Rheumatoid Arthritis Related Conditions | 0.486 | 0.027 | 0.161 | [213] |
| Ulcerative Colitis | 0.471 | 0.033 | 0.263 | [172] |
| Alopecia | 0.321 | 0.023 | 0.39 | [1] |
| Gout Related Crystal Arthropathies | 0.49 | 0.023 | 0.41 | [45] |
| ADHD | 0.763 | 0.009 | 0.257 | [202] |
| Autism | 0.924 | 0.006 | 0.462 | [236] |
| Bipolar Disorder | 0.676 | 0.011 | 0.432 | [230] |

Continued on next page

**Table A.6 – continued from previous page**

| Disease | Family $h^2$ Adjusted | Family $h^2$ SD | GWAS $h^2$ | GWAS Reference |
|---|---|---|---|---|
| Depression | 0.579 | 0.006 | 0.175 | [185] |
| Eating Disorder | 0.569 | 0.031 | 0.559 | [325] |
| Migraine | 0.374 | 0.01 | 0.146 | [101] |
| OCD | 0.657 | 0.023 | 0.14 | [61] |
| PTSD | 0.577 | 0.021 | 0.062 | [272] |
| Personality Disorder | 0.52 | 0.028 | 0.55 | [285] |
| Schizophrenia Related Psychosis | 0.562 | 0.022 | 0.454 | [255] |
| Glaucoma | 0.698 | 0.008 | 0.06 | [268] |
| Refraction Accommodation Disorders | 0.611 | 0.015 | 0.12 | [75] |
| Hearing Loss | 0.217 | 0.012 | 0.05 | [163] |
| Emphysema COPD | 0.534 | 0.019 | 0.377 | [326] |
| Urinary Calculus | 0.357 | 0.015 | 0.048 | [211] |

| | 15-17 | 18-24 | 25+ | 15+ |
|---|---|---|---|---|
| Total | 11,410,194 | 13,649,432 | 10,392,677 | 35,452,303 |
| Biological | 10,301,224 | 12,434,801 | 9,680,661 | 32,416,686 |
| Adopted | 303,205 | 325,272 | 220,020 | 848,497 |
| Step | 805,765 | 889,359 | 491,996 | 2,187,120 |
| Biological % | 90.3% | 91.1% | 93.1% | 91.4% |

**Table A.7:** Children by household type and age from U.S. Census 2010

| Household | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| Two parents | 52,153 | 51,785 | 51,835 | 51,823 | 51,456 |
| Both biological | 46,681 | 46,427 | 46,364 | 46,438 | 46,405 |
| Biological % | 89.5% | 89.7% | 89.4% | 89.6% | 90.2% |

**Table A.8:** Children under 18 by household type from Current Population Survey 2007-2011

# REFERENCES


[1] Kaustubh Adhikari, Tania Fontanil, Santiago Cal, Javier Mendoza-Revilla, Macarena Fuentes-Guajardo, Juan-Camilo Chacón-Duque, Farah Al-Saadi, Jeanette A. Johansson, Mirsha Quinto-Sanchez, Victor Acuña-Alonzo, Claudia Jaramillo, William Arias, Rodrigo Barquera Lozano, Gastón Macín Pérez, Jorge Gómez-Valdés, Hugo Villamil-Ramírez, Tábita Hunemeier, Virginia Ramallo, Caio C. Silva de Cerqueira, Malena Hurtado, Valeria Villegas, Vanessa Granja, Carla Gallo, Giovanni Poletti, Lavinia Schuler-Faccini, Francisco M. Salzano, Maria-Cátira Bortolini, Samuel Canizales-Quinteros, Francisco Rothhammer, Gabriel Bedoya, Rolando Gonzalez-José, Denis Headon, Carlos López-Otín, Desmond J. Tobin, David Balding, and Andrés Ruiz-Linares. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nature Communications*, 7:10815, March 2016.

[2] Ali Akyol, Banu Bicerol, Sema Ertug, Hatice Ertabaklar, and Nefati Kiylioglu. Epilepsy and seropositivity rates of Toxocara canis and Toxoplasma gondii. *Seizure*, 16(3):233–237, April 2007.

[3] Nely Aldrich, Meg Gerstenblith, Pingfu Fu, Marie S. Tuttle, Priya Varma, Erica Gotow, Kevin D. Cooper, Margaret Mann, and Daniel L. Popkin. Genetic vs Environmental Factors That Correlate With Rosacea: A Cohort-Based Survey of Twins. *JAMA Dermatology*, 151(11):1213–1219, November 2015.

[4] Eileen S. Alexander, Lisa J. Martin, Margaret H. Collins, Leah C. Kottyan, Heidi Sucharew, Hua He, Vincent A. Mukkada, Paul A. Succop, J. Pablo Abonia, Heather Foote, Michael D. Eby, Tommie M. Grotjan, Alexandria J. Greenler, Evan S. Dellon, Jeffrey G. Demain, Glenn T. Furuta, Larry E. Gurian, John B. Harley, Russell J. Hopp, Amir Kagalwalla, Ajay Kaul, Kari C. Nadeau, Richard J. Noel, Philip E. Putnam, Karl F. von Tiehl, and Marc E. Rothenberg. Twin and family studies reveal strong environmental and weaker genetic cues explaining heritability of eosinophilic esophagitis. *Journal of Allergy and Clinical Immunology*, 134(5):1084 – 1092.e1, 2014.

[5] Sophia Ananiadou, Paul Thompson, Raheel Nawaz, John McNaught, and Douglas B. Kell. Event-based text mining for biology and functional genomics. *Briefings in Functional Genomics*, 14(3):213–230, May 2015.

[6] Margaret Jean Anderson. *Carl Linnaeus: Father of Classification*. Enslow Pub Inc, Springfield, NJ, August 1997.

[7] Verneri Anttila, Brendan Bulik-Sullivan, Hilary Kiyo Finucane, and et al. Analysis of shared heritability in common disorders of the brain. *bioRxiv*, 2016.

[8] J. Fernando Arevalo, Rubens Belfort, Cristina Muccioli, and Juan V. Espinoza. Ocular toxoplasmosis in the developing world. *International Ophthalmology Clinics*, 50(2):57–69, 2010.

[9] Yoav Arnson, Howard Amital, and Yehuda Shoenfeld. Vitamin D and autoimmunity: New aetiological and therapeutic considerations. *Annals of the Rheumatic Diseases*, 66(9):1137–1142, September 2007.

[10] Alan R Aronson and François-Michel Lang. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–236, 2010.

[11] Autism and Developmental Disabilities Monitoring Network Surveillance Year 2002 Principal Investigators and Centers for Disease Control and Prevention. Prevalence of autism spectrum disorders–autism and developmental disabilities monitoring network, 14 sites, United States, 2002. *Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002)*, 56(1):12–28, February 2007.

[12] Claus Bachert, Ruby Pawankar, Luo Zhang, Chaweewan Bunnag, Wytske J. Fokkens, Daniel L. Hamilos, Orathai Jirapongsananuruk, Robert Kern, Eli O. Meltzer, Joaquim Mullol, Robert Naclerio, Renata Pilan, Chae-Seo Rhee, Harumi Suzaki, Richard Voegels, and Michael Blaiss. ICON: Chronic rhinosinusitis. *World Allergy Organization Journal*, 7(1):1–28, 2014.

[13] Peter Bailey, David K. Chang, Katia Nones, Amber L. Johns, Ann-Marie Patch, Marie-Claude Gingras, David K. Miller, Angelika N. Christ, Tim J. C. Bruxner, Michael C. Quinn, Craig Nourse, L. Charles Murtaugh, Ivon Harliwong, Senel Idrisoglu, Suzanne Manning, Ehsan Nourbakhsh, Shivangi Wani, Lynn Fink, Oliver Holmes, Venessa Chin, Matthew J. Anderson, Stephen Kazakoff, Conrad Leonard, Felicity Newell, Nick Waddell, Scott Wood, Qinying Xu, Peter J. Wilson, Nicole Cloonan, Karin S. Kassahn, Darrin Taylor, Kelly Quek, Alan Robertson, Lorena Pantano, Laura Mincarelli, Luis N. Sanchez, Lisa Evers, Jianmin Wu, Mark Pinese, Mark J. Cowley, Marc D. Jones, Emily K. Colvin, Adnan M. Nagrial, Emily S. Humphrey, Lorraine A. Chantrill, Amanda Mawson, Jeremy Humphris, Angela Chou, Marina Pajic, Christopher J. Scarlett, Andreia V. Pinho, Marc Giry-Laterriere, Ilse Rooman, Jaswinder S. Samra, James G. Kench, Jessica A. Lovell, Neil D. Merrett, Christopher W. Toon, Krishna Epari, Nam Q. Nguyen, Andrew Barbour, Nikolajs Zeps, Kim Moran-Jones, Nigel B. Jamieson, Janet S. Graham, Fraser Duthie, Karin Oien, Jane Hair, Robert Grützmann, Anirban Maitra, Christine A. Iacobuzio-Donahue, Christopher L. Wolfgang, Richard A. Morgan, Rita T. Lawlor, Vincenzo Corbo, Claudio Bassi, Borislav Rusev, Paola Capelli, Roberto Salvia, Giampaolo Tortora, Debabrata Mukhopadhyay, Gloria M. Petersen, Australian Pancreatic Cancer Genome Initiative, Donna M. Munzy, William E. Fisher, Saadia A. Karim, James R. Eshleman, Ralph H. Hruban, Christian Pilarsky, Jennifer P. Morton, Owen J. Sansom, Aldo Scarpa, Elizabeth A. Musgrove, Ulla-Maja Hagbo Bailey, Oliver Hofmann, Robert L. Sutherland, David A. Wheeler, Anthony J. Gill, Richard A. Gibbs, John V. Pearson, Nicola Waddell, Andrew V. Biankin, and Sean M. Grimmond. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52, March 2016.

[14] D. J. P. Barker, J. G. Eriksson, T. Forsén, and C. Osmond. Fetal origins of adult

disease: Strength of effects and biological basis. *International Journal of Epidemiology*, 31(6):1235–1239, January 2002.

[15] J Barnard, R McCulloch, and XL Meng. "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage". *Statistica Sinica*, 10(4):1281–1311, 2000.

[16] Laurence S. Baskin, editor. *Hypospadias and Genital Development*. Springer, Place of publication not identified, softcover reprint of the original 1st ed. 2004 edition edition, October 2013.

[17] V. Bataille, H. Snieder, A. J. MacGregor, P. Sasieni, and T. D. Spector. The influence of genetics and environmental factors in the pathogenesis of acne: A twin study of acne in women. *The Journal of Investigative Dermatology*, 119(6):1317–1322, December 2002.

[18] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models using lme4. *arXiv:1406.5823 [stat]*, June 2014.

[19] Michele C. Battié, Alfredo Ortega-Alonso, Riikka Niemelainen, Kevin Gill, Esko Levalahti, Tapio Videman, and Jaakko Kaprio. Lumbar spinal stenosis is a highly genetic condition partly mediated by disc degeneration. *Arthritis & Rheumatology (Hoboken, N.J.)*, 66(12):3505–3510, December 2014.

[20] Tracy Ann Becerra, Michelle Wilhelm, Jørn Olsen, Myles Cockburn, and Beate Ritz. Ambient air pollution and autism in Los Angeles county, California. *Environmental Health Perspectives*, 121(3):380–386, March 2013.

[21] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300, 1995.

[22] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300, 1995. Qe453 Times Cited:3303 Cited References Count:14.

[23] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. 504BV Times Cited:402 Cited References Count:43.

[24] Camillo Bérénos, Philip A. Ellis, Jill G. Pilkington, and Josephine M. Pemberton. Estimating quantitative genetic parameters in wild populations: A comparison of pedigree and genomic approaches. *Molecular Ecology*, 23(14):3434–3451, 2014.

[25] Jamee M. Berg and Daniel H. Geschwind. Autism genetics: Searching for specificity and convergence. *Genome Biology*, 13(7):247, July 2012.

[26] David R. Blair, Kanix Wang, Svetlozar Nestorov, James A. Evans, and Andrey Rzhetsky. Quantifying the Impact and Extent of Undocumented Biomedical Synonymy. *PLOS Computational Biology*, 10(9):e1003799, September 2014.

[27] N. Blanchard, I. R. Dunay, and D. Schlüter. Persistence of Toxoplasma gondii in the central nervous system: A fine-tuned balance between the parasite, the brain and the immune system. *Parasite Immunology*, 37(3):150–158, March 2015.

[28] Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270, January 2004.

[29] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186, June 2017.

[30] Robert L. Brent. Environmental causes of human congenital malformations: The pediatrician's role in dealing with these complex clinical problems caused by a multiplicity of environmental and genetic factors. *Pediatrics*, 113(4 Suppl):957–968, April 2004.

[31] Alan S. Brown, Catherine A. Schaefer, Charles P. Quesenberry, Liyan Liu, Vicki P. Babulas, and Ezra S. Susser. Maternal exposure to toxoplasmosis and risk of schizophrenia in adult offspring. *The American Journal of Psychiatry*, 162(4):767–773, April 2005.

[32] M. A. Brown, L. G. Kennedy, A. J. MacGregor, C. Darke, E. Duncan, J. L. Shatford, A. Taylor, A. Calin, and P. Wordsworth. Susceptibility to ankylosing spondylitis in twins: The role of genes, HLA, and the environment. *Arthritis and Rheumatism*, 40(10):1823–1828, October 1997.

[33] Germaine M. Buck Louis, Melissa M. Smarr, and Chirag J. Patel. The Exposome Research Paradigm: An Opportunity to Understand the Environmental Basis for Human Health and Disease. *Current Environmental Health Reports*, 4(1):89–98, March 2017.

[34] Cynthia M. Bulik, Patrick F. Sullivan, Federica Tozzi, Helena Furberg, Paul Lichtenstein, and Nancy L. Pedersen. Prevalence, heritability, and prospective risk factors for anorexia nervosa. *Archives of General Psychiatry*, 63(3):305–312, March 2006.

[35] B Bulik-Sullivan, HK Finucane, V Anttila, and et al. An Atlas of Genetic Correlations across Human Diseases and Traits. *Nature genetics*, 47(11):1236–1241, 2015.

[36] Brendan K. Bulik-Sullivan, Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, March 2015.

[37] CDC. Toxoplasmosis. `https://www.cdc.gov/parasites/toxoplasmosis/`, 2017.

[38] U.S. CDC. Centers for Disease Control and Prevention: Health Insurance. `https://www.cdc.gov/nchs/fastats/health-insurance.htm`.

[39] Patrícia B. S. Celestino-Soper, Sara Violante, Emily L. Crawford, Rui Luo, Anath C. Lionel, Elsa Delaby, Guiqing Cai, Bekim Sadikovic, Kwanghyuk Lee, Charlene Lo, Kun Gao, Richard E. Person, Timothy J. Moss, Jennifer R. German, Ni Huang, Marwan Shinawi, Diane Treadwell-Deering, Peter Szatmari, Wendy Roberts, Bridget Fernandez, Richard J. Schroer, Roger E. Stevenson, Joseph D. Buxbaum, Catalina Betancur, Stephen W. Scherer, Stephan J. Sanders, Daniel H. Geschwind, James S. Sutcliffe, Matthew E. Hurles, Ronald J. A. Wanders, Chad A. Shaw, Suzanne M. Leal, Edwin H. Cook, Robin P. Goin-Kochel, Frédéric M. Vaz, and Arthur L. Beaudet. A common X-linked inborn error of carnitine biosynthesis may be a risk factor for nondysmorphic autism. *Proceedings of the National Academy of Sciences of the United States of America*, 109(21):7974–7981, May 2012.

[40] Tyani D. Chan, Katherine Wood, Jana R. Hermes, Danyal Butt, Christopher J. Jolly, Antony Basten, and Robert Brink. Elimination of Germinal-Center-Derived Self-Reactive B Cells Is Governed by the Location and Concentration of Self-Antigen. *Immunity*, 37(5):893–904, November 2012.

[41] Anne Charmantier and Denis Réale. How do misassigned paternities affect the estimation of heritability in the wild? *Molecular Ecology*, 14(9):2839–2850, 2005.

[42] Guo-Bo Chen, Sang Hong Lee, Marie-Jo A. Brion, Grant W. Montgomery, Naomi R. Wray, Graham L. Radford-Smith, and Peter M. Visscher. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Human Molecular Genetics*, 23(17):4710–4720, September 2014.

[43] Tzu-Ying Chiang, Tzu-Hsuen Yuan, Ruei-Hao Shie, Chen-Fang Chen, and Chang-Chuan Chan. Increased incidence of allergic rhinitis, bronchitis and asthma, in children living near a petrochemical complex with SO2 pollution. *Environment International*, 96:1–7, November 2016.

[44] S. Chib and E. Greenberg. "Analysis of multivariate probit models". *Biometrika*, 85(2):347–361, 1998.

[45] Geetha Chittoor, Karin Haack, Nitesh R. Mehta, Sandra Laston, Shelley A. Cole, Anthony G. Comuzzie, Nancy F. Butte, and V. Saroja Voruganti. Genetic variation underlying renal uric acid excretion in Hispanic children: The Viva La Familia Study. *BMC Medical Genetics*, 18, January 2017.

[46] Hyun-Ju Cho, Mee Hyun Song, Soo-Young Choi, Jeongho Kim, Jinwook Lee, Un-Kyung Kim, Jinwoong Bok, and Jae Young Choi. Genetic analysis of the CHD7 gene in Korean patients with CHARGE syndrome. *Gene*, 517(2):164–168, April 2013.

[47] The International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–998, April 2010.

[48] Colin R. Cooke and Theodore J. Iwashyna. Using existing data to address important clinical questions in critical care. *Critical Care Medicine*, 41(3):886–896, March 2013.

[49] Margaret M. Cortese, Jacqueline E. Tate, Lone Simonsen, Laurel Edelman, and Umesh D. Parashar. Reduction in gastroenteritis in United States children and correlation with early rotavirus vaccine uptake from national medical claims databases. *The Pediatric Infectious Disease Journal*, 29(6):489–494, June 2010.

[50] D. R. Cox and E. J. Snell. *Analysis of Binary Data*. London: Chapman and Hall, 1989.

[51] Lisa A. Croen, Judith K. Grether, Jenny Hoogstrate, and Steve Selvin. The changing prevalence of autism in California. *Journal of Autism and Developmental Disorders*, 32(3):207–215, June 2002.

[52] Cross Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature genetics*, 45(9):984–994, 2013.

[53] William Crowe, Philip J. Allsopp, Gene E. Watson, Pamela J. Magee, J. J. Strain, David J. Armstrong, Elizabeth Ball, and Emeir M. McSorley. Mercury as an environmental stimulus in the development of autoimmunity - A systematic review. *Autoimmunity Reviews*, September 2016.

[54] Kamila Czene, Paul Lichtenstein, and Kari Hemminki. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family-cancer database. *International Journal of Cancer*, 99(2):260–266, May 2002.

[55] Jesse M. Damsker, Anna M. Hansen, and Rachel R. Caspi. Th1 and Th17 cells: Adversaries and collaborators. *Annals of the New York Academy of Sciences*, 1183:211–221, January 2010.

[56] Lea K. Davis, Dongmei Yu, Clare L. Keenan, Eric R. Gamazon, Anuar I. Konkashbaev, Eske M. Derks, Benjamin M. Neale, Jian Yang, S. Hong Lee, Patrick Evans, Cathy L. Barr, Laura Bellodi, Fortu Benarroch, Gabriel Bedoya Berrio, Oscar J. Bienvenu, Michael H. Bloch, Rianne M. Blom, Ruth D. Bruun, Cathy L. Budman, Beatriz Camarena, Desmond Campbell, Carolina Cappi, Julio C. Cardona Silgado, Danielle C. Cath, Maria C. Cavallini, Denise A. Chavira, Sylvain Chouinard, David V. Conti, Edwin H. Cook, Vladimir Coric, Bernadette A. Cullen, Dieter Deforce, Richard Delorme, Yves Dion, Christopher K. Edlund, Karin Egberts, Peter Falkai, Thomas V. Fernandez, Patience J. Gallagher, Helena Garrido, Daniel Geller, Simon L. Girard, Hans J. Grabe, Marco A. Grados, Benjamin D. Greenberg, Varda Gross-Tsur, Stephen Haddad, Gary A. Heiman, Sian M. J. Hemmings, Ana G. Hounie, Cornelia Illmann, Joseph Jankovic, Michael A. Jenike, James L. Kennedy, Robert A. King, Barbara Kremeyer, Roger Kurlan, Nuria Lanzagorta, Marion Leboyer, James F. Leckman, Leonhard Lennertz, Chunyu Liu, Christine Lochner, Thomas L. Lowe, Fabio Macciardi, James T. McCracken, Lauren M. McGrath, Sandra C. Mesa Restrepo, Rainald Moessner, Jubel

Morgan, Heike Muller, Dennis L. Murphy, Allan L. Naarden, William Cornejo Ochoa, Roel A. Ophoff, Lisa Osiecki, Andrew J. Pakstis, Michele T. Pato, Carlos N. Pato, John Piacentini, Christopher Pittenger, Yehuda Pollak, Scott L. Rauch, Tobias J. Renner, Victor I. Reus, Margaret A. Richter, Mark A. Riddle, Mary M. Robertson, Roxana Romero, Maria C. Rosàrio, David Rosenberg, Guy A. Rouleau, Stephan Ruhrmann, Andres Ruiz-Linares, Aline S. Sampaio, Jack Samuels, Paul Sandor, Brooke Sheppard, Harvey S. Singer, Jan H. Smit, Dan J. Stein, E. Strengman, Jay A. Tischfield, Ana V. Valencia Duarte, Homero Vallada, Filip Van Nieuwerburgh, Jeremy Veenstra-VanderWeele, Susanne Walitza, Ying Wang, Jens R. Wendland, Herman G. M. Westenberg, Yin Yao Shugart, Euripedes C. Miguel, William McMahon, Michael Wagner, Humberto Nicolini, Danielle Posthuma, Gregory L. Hanna, Peter Heutink, Damiaan Denys, Paul D. Arnold, Ben A. Oostra, Gerald Nestadt, Nelson B. Freimer, David L. Pauls, Naomi R. Wray, S. Evelyn Stewart, Carol A. Mathews, James A. Knowles, Nancy J. Cox, and Jeremiah M. Scharf. Partitioning the Heritability of Tourette Syndrome and Obsessive Compulsive Disorder Reveals Differences in Genetic Architecture. *PLOS Genetics*, 9(10):e1003864, October 2013.

[57] Antoine Laurent de Jussieu. *Antonii Laurentii de Jussieu Genera Plantarum :Secundum Ordines Naturales Disposita, Juxta Methodum in Horto Regio Parisiensi Exaratam, Anno M.DCC.LXXIV.* apud viduam Herissant et Theophilum Barrois,, Parisiis :, 1789.

[58] Bert De Smedt, Ann Swillen, Lieven Verschaffel, and Pol Ghesquière. Mathematical learning disabilities in children with 22q11.2 deletion syndrome: A review. *Developmental Disabilities Research Reviews*, 15(1):4–10, 2009.

[59] Pierre de Villemereuil, Olivier Gimenez, and Blandine Doligez. Comparing parent–offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: A simulation study for Gaussian and binary traits. *Methods in Ecology and Evolution*, 4(3):260–275, March 2013.

[60] M. Dedicoat and N. Livesley. Management of toxoplasmic encephalitis in HIV-infected adults (with an emphasis on resource-poor settings). *The Cochrane Database of Systematic Reviews*, (3):CD005420, July 2006.

[61] A. den Braber, N. R. Zilhão, I. O. Fedko, J.-J. Hottenga, R. Pool, D. J. A. Smit, D. C. Cath, and D. I. Boomsma. Obsessive-compulsive symptoms in a large population-based twin-family sample are predicted by clinically based polygenic scores and by genome-wide SNPs. *Translational Psychiatry*, 6:e731, February 2016.

[62] Joshua C. Denny, Lisa Bastarache, Marylyn D. Ritchie, Robert J. Carroll, Raquel Zink, Jonathan D. Mosley, Julie R. Field, Jill M. Pulley, Andrea H. Ramirez, Erica Bowton, Melissa A. Basford, David S. Carrell, Peggy L. Peissig, Abel N. Kho, Jennifer A. Pacheco, Luke V. Rasmussen, David R. Crosslin, Paul K. Crane, Jyotishman Pathak, Suzette J. Bielinski, Sarah A. Pendergrass, Hua Xu, Lucia A. Hindorff, Rongling Li, Teri A. Manolio, Christopher G. Chute, Rex L. Chisholm, Eric B. Larson, Gail P. Jarvik, Murray H. Brilliant, Catherine A. McCarty, Iftikhar J. Kullo, Jonathan L.

Haines, Dana C. Crawford, Daniel R. Masys, and Dan M. Roden. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31(12):1102–1111, December 2013.

[63] Joshua C. Denny, Marylyn D. Ritchie, Melissa A. Basford, Jill M. Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R. Masys, Dan M. Roden, and Dana C. Crawford. PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics (Oxford, England)*, 26(9):1205–1210, May 2010.

[64] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *Journal of biomedical informatics*, 47:1–10, February 2014.

[65] Richard Doll, Richard Peto, Jillian Boreham, and Isabelle Sutherland. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ*, 328(7455):1519, June 2004.

[66] Bradley Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, Pa, January 1987.

[67] Bradley Efron. The Bootstrap and Markov-Chain Monte Carlo. *Journal of Biopharmaceutical Statistics*, 21(6):1052–1062, November 2011.

[68] A. Eichhorn, S. Lochner, and G. G. Belz. [Vitamin D for prevention of diseases?]. *Deutsche Medizinische Wochenschrift (1946)*, 137(17):906–912, April 2012.

[69] Jay W. Ellison, Jill A. Rosenfeld, and Lisa G. Shaffer. Genetic basis of intellectual disability. *Annual Review of Medicine*, 64:441–450, 2013.

[70] István László Endlicher. *Genera Plantarum Secundum Ordines Naturales Disposita.* F. Beck, Vindobonae, 1836-1840.

[71] A M Ercolini and S D Miller. The role of infections in autoimmune disease. *Clinical and Experimental Immunology*, 155(1):1–15, January 2009.

[72] J. G. Eriksson, T. Forsén, J. Tuomilehto, C. Osmond, and D. J. P. Barker. Early growth and coronary heart disease in later life: Longitudinal study. *BMJ*, 322(7292):949–953, April 2001.

[73] D. Falconer and T. Mackay. *Introduction to Quantitative Genetics 4th Edn.* Harlow, UK: Longman Scientific and Technical, 1996.

[74] D. S. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76, 1965.

[75] Qiao Fan, Virginie J. M. Verhoeven, Robert Wojciechowski, Veluchamy A. Barathi, Pirro G. Hysi, Jeremy A. Guggenheim, René Höhn, Veronique Vitart, Anthony P. Khawaja, Kenji Yamashiro, S Mohsen Hosseini, Terho Lehtimäki, Yi Lu, Toomas

Haller, Jing Xie, Cécile Delcourt, Mario Pirastu, Juho Wedenoja, Puya Gharahkhani, Cristina Venturini, Masahiro Miyake, Alex W. Hewitt, Xiaobo Guo, Johanna Mazur, Jenifer E. Huffman, Katie M. Williams, Ozren Polasek, Harry Campbell, Igor Rudan, Zoran Vatavuk, James F. Wilson, Peter K. Joshi, George McMahon, Beate St Pourcain, David M. Evans, Claire L. Simpson, Tae-Hwi Schwantes-An, Robert P. Igo, Alireza Mirshahi, Audrey Cougnard-Gregoire, Céline Bellenguez, Maria Blettner, Olli Raitakari, Mika Kähönen, Ilkka Seppala, Tanja Zeller, Thomas Meitinger, Janina S. Ried, Christian Gieger, Laura Portas, Elisabeth M. van Leeuwen, Najaf Amin, André G. Uitterlinden, Fernando Rivadeneira, Albert Hofman, Johannes R. Vingerling, Ya Xing Wang, Xu Wang, Eileen Tai-Hui Boh, M. Kamran Ikram, Charumathi Sabanayagam, Preeti Gupta, Vincent Tan, Lei Zhou, Candice E. H. Ho, Wan'e Lim, Roger W. Beuerman, Rosalynn Siantar, E-Shyong Tai, Eranga Vithana, Evelin Mihailov, Chiea-Chuen Khor, Caroline Hayward, Robert N. Luben, Paul J. Foster, Barbara E. K. Klein, Ronald Klein, Hoi-Suen Wong, Paul Mitchell, Andres Metspalu, Tin Aung, Terri L. Young, Mingguang He, Olavi Pärssinen, Cornelia M. van Duijn, Jie Jin Wang, Cathy Williams, Jost B. Jonas, Yik-Ying Teo, David A. Mackey, Konrad Oexle, Nagahisa Yoshimura, Andrew D. Paterson, Norbert Pfeiffer, Tien-Yin Wong, Paul N. Baird, Dwight Stambolian, Joan E. Bailey Wilson, Ching-Yu Cheng, Christopher J. Hammond, Caroline C. W. Klaver, Seang-Mei Saw, Jugnoo S. Rahi, Jean-François Korobelnik, John P. Kemp, Nicholas J. Timpson, George Davey Smith, Jamie E. Craig, Kathryn P. Burdon, Rhys D. Fogarty, Sudha K. Iyengar, Emily Chew, Sarayut Janmahasatian, Nicholas G. Martin, Stuart MacGregor, Liang Xu, Maria Schache, Vinay Nangia, Songhomitra Panda-Jonas, Alan F. Wright, Jeremy R. Fondran, Jonathan H. Lass, Sheng Feng, Jing Hua Zhao, Kay-Tee Khaw, Nick J. Wareham, Taina Rantanen, Jaakko Kaprio, Chi Pui Pang, Li Jia Chen, Pancy O. Tam, Vishal Jhanji, Alvin L. Young, Angela Döring, Leslie J. Raffel, Mary-Frances Cotch, Xiaohui Li, Shea Ping Yip, Maurice K.H. Yap, Ginevra Biino, Simona Vaccargiu, Maurizio Fossarello, Brian Fleck, Seyhan Yazar, Jan Willem L. Tideman, Milly Tedja, Margaret M. Deangelis, Margaux Morrison, Lindsay Farrer, Xiangtian Zhou, Wei Chen, Nobuhisa Mizuki, Akira Meguro, and Kari Matti Mäkelä. Meta-analysis of gene–environment-wide association scans accounting for education level identifies additional loci for refractive error. *Nature Communications*, 7, March 2016.

[76] Zuoxu Fan, Yaoyao Wu, Jian Shen, Tao Ji, and Renya Zhan. Schizophrenia and the risk of cardiovascular diseases: A meta-analysis of thirteen cohort studies. *Journal of Psychiatric Research*, 47(11):1549–1556, November 2013.

[77] Gordon Fehringer, Peter Kraft, Paul D. Pharoah, Rosalind A. Eeles, Nilanjan Chatterjee, Fredrick R. Schumacher, Joellen M. Schildkraut, Sara Lindström, Paul Brennan, Heike Bickeböller, Richard S. Houlston, Maria Teresa Landi, Neil Caporaso, Angela Risch, Ali Amin Al Olama, Sonja I. Berndt, Edward L. Giovannucci, Henrik Grönberg, Zsofia Kote-Jarai, Jing Ma, Kenneth Muir, Meir J. Stampfer, Victoria L. Stevens, Fredrik Wiklund, Walter C. Willett, Ellen L. Goode, Jennifer B. Permuth, Harvey A. Risch, Brett M. Reid, Stephane Bezieau, Hermann Brenner, Andrew T. Chan, Jenny Chang-Claude, Thomas J. Hudson, Jonathan K. Kocarnik, Polly A. Newcomb,

Robert E. Schoen, Martha L. Slattery, Emily White, Muriel A. Adank, Habibul Ahsan, Kristiina Aittomäki, Laura Baglietto, Carl Blomquist, Federico Canzian, Kamila Czene, Isabel Dos-Santos-Silva, A. Heather Eliassen, Jonine D. Figueroa, Dieter Flesch-Janys, Olivia Fletcher, Montserrat Garcia-Closas, Mia M. Gaudet, Nichola Johnson, Per Hall, Aditi Hazra, Rebecca Hein, Albert Hofman, John L. Hopper, Astrid Irwanto, Mattias Johansson, Rudolf Kaaks, Muhammad G. Kibriya, Peter Lichtner, Jianjun Liu, Eiliv Lund, Enes Makalic, Alfons Meindl, Bertram Müller-Myhsok, Taru A. Muranen, Heli Nevanlinna, Petra H. Peeters, Julian Peto, Ross L. Prentice, Nazneen Rahman, Maria Jose Sanchez, Daniel F. Schmidt, Rita K. Schmutzler, Melissa C. Southey, Rulla Tamimi, Ruth C. Travis, Clare Turnbull, Andre G. Uitterlinden, Zhaoming Wang, Alice S. Whittemore, Xiaohong R. Yang, Wei Zheng, Daniel D. Buchanan, Graham Casey, David V. Conti, Christopher K. Edlund, Steven Gallinger, Robert W. Haile, Mark Jenkins, Loïc Le Marchand, Li Li, Noralene M. Lindor, Stephanie L. Schmit, Stephen N. Thibodeau, Michael O. Woods, Thorunn Rafnar, Julius Gudmundsson, Simon N. Stacey, Kari Stefansson, Patrick Sulem, Y. Ann Chen, Jonathan P. Tyrer, David C. Christiani, Yongyue Wei, Hongbing Shen, Zhibin Hu, Xiao-Ou Shu, Kouya Shiraishi, Atsushi Takahashi, Yohan Bossé, Ma'en Obeidat, David Nickle, Wim Timens, Matthew L. Freedman, Qiyuan Li, Daniela Seminara, Stephen J. Chanock, Jian Gong, Ulrike Peters, Stephen B. Gruber, Christopher I. Amos, Thomas A. Sellers, Douglas F. Easton, David J. Hunter, Christopher A. Haiman, Brian E. Henderson, Rayjean J. Hung, Ovarian Cancer Association Consortium (OCAC), PRACTICAL Consortium, Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON), Colorectal Transdisciplinary (CORECT) Study, and African American Breast Cancer Consortium (AABC) and African Ancestry Prostate Cancer Consortium (AAPC). Cross-Cancer Genome-Wide Analysis of Lung, Ovary, Breast, Prostate, and Colorectal Cancer Reveals Novel Pleiotropic Associations. *Cancer Research*, 76(17):5103–5114, September 2016.

[78] A. R. Feinstein. THE PRE-THERAPEUTIC CLASSIFICATION OF COMORBIDITY IN CHRONIC DISEASE. *Journal of Chronic Diseases*, 23(7):455–468, December 1970.

[79] Joseph Felsenstein. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4):783–791, 1985.

[80] R.A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[81] Ronald Aylmer Fisher. 009: The Correlation Between Relatives on the Supposition of Mendelian Inheritance. 1918.

[82] Elaine W. Flagg, Robert Schwartz, and Hillard Weinstock. Prevalence of anogenital warts among participants in private health plans in the United States, 2003-2010: Potential impact of human papillomavirus vaccination. *American Journal of Public Health*, 103(8):1428–1435, August 2013.

[83] Andre Franke, Dermot P. B. McGovern, Jeffrey C. Barrett, Kai Wang, Graham L. Radford-Smith, Tariq Ahmad, Charlie W. Lees, Tobias Balschun, James Lee, Rebecca Roberts, Carl A. Anderson, Joshua C. Bis, Suzanne Bumpstead, David Ellinghaus, Eleonora M. Festen, Michel Georges, Todd Green, Talin Haritunians, Luke Jostins, Anna Latiano, Christopher G. Mathew, Grant W. Montgomery, Natalie J. Prescott, Soumya Raychaudhuri, Jerome I. Rotter, Philip Schumm, Yashoda Sharma, Lisa A. Simms, Kent D. Taylor, David Whiteman, Cisca Wijmenga, Robert N. Baldassano, Murray Barclay, Theodore M. Bayless, Stephan Brand, Carsten Büning, Albert Cohen, Jean-Frederick Colombel, Mario Cottone, Laura Stronati, Ted Denson, Martine De Vos, Renata D'Inca, Marla Dubinsky, Cathryn Edwards, Tim Florin, Denis Franchimont, Richard Gearry, Jürgen Glas, Andre Van Gossum, Stephen L. Guthery, Jonas Halfvarson, Hein W. Verspaget, Jean-Pierre Hugot, Amir Karban, Debby Laukens, Ian Lawrance, Marc Lemann, Arie Levine, Cecile Libioulle, Edouard Louis, Craig Mowat, William Newman, Julián Panés, Anne Phillips, Deborah D. Proctor, Miguel Regueiro, Richard Russell, Paul Rutgeerts, Jeremy Sanderson, Miquel Sans, Frank Seibold, A. Hillary Steinhart, Pieter C. F. Stokkers, Leif Torkvist, Gerd Kullak-Ublick, David Wilson, Thomas Walters, Stephan R. Targan, Steven R. Brant, John D. Rioux, Mauro D'Amato, Rinse K. Weersma, Subra Kugathasan, Anne M. Griffiths, John C. Mansfield, Severine Vermeire, Richard H. Duerr, Mark S. Silverberg, Jack Satsangi, Stefan Schreiber, Judy H. Cho, Vito Annese, Hakon Hakonarson, Mark J. Daly, and Miles Parkes. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, 42(12):1118–1125, December 2010.

[84] Florian Friedmacher and Prem Puri. Hirschsprung's disease associated with Down syndrome: A meta-analysis of incidence, functional outcomes and mortality. *Pediatric Surgery International*, 29(9):937–946, September 2013.

[85] Lars G. Fritsche, Wilmar Igl, Jessica N. Cooke Bailey, Felix Grassmann, Sebanti Sengupta, Jennifer L. Bragg-Gresham, Kathryn P. Burdon, Scott J. Hebbring, Cindy Wen, Mathias Gorski, Ivana K. Kim, David Cho, Donald Zack, Eric Souied, Hendrik P. N. Scholl, Elisa Bala, Kristine E. Lee, David J. Hunter, Rebecca J. Sardell, Paul Mitchell, Joanna E. Merriam, Valentina Cipriani, Joshua D. Hoffman, Tina Schick, Yara T. E. Lechanteur, Robyn H. Guymer, Matthew P. Johnson, Yingda Jiang, Chloe M. Stanton, Gabriëlle H. S. Buitendijk, Xiaowei Zhan, Alan M. Kwong, Alexis Boleda, Matthew Brooks, Linn Gieser, Rinki Ratnapriya, Kari E. Branham, Johanna R. Foerster, John R. Heckenlively, Mohammad I. Othman, Brendan J. Vote, Helena Hai Liang, Emmanuelle Souzeau, Ian L. McAllister, Timothy Isaacs, Janette Hall, Stewart Lake, David A. Mackey, Ian J. Constable, Jamie E. Craig, Terrie E. Kitchner, Zhenglin Yang, Zhiguang Su, Hongrong Luo, Daniel Chen, Hong Ouyang, Ken Flagg, Danni Lin, Guanping Mao, Henry Ferreyra, Klaus Stark, Claudia N. von Strachwitz, Armin Wolf, Caroline Brandl, Guenther Rudolph, Matthias Olden, Margaux A. Morrison, Denise J. Morgan, Matthew Schu, Jeeyun Ahn, Giuliana Silvestri, Evangelia E. Tsironi, Kyu Hyung Park, Lindsay A. Farrer, Anton Orlin, Alexander Brucker, Mingyao Li, Christine A. Curcio, Saddek Mohand-Saïd, José-Alain Sahel, Isabelle Audo, Mustapha Benchaboune, Angela J. Cree, Christina A. Rennie, Srinivas V. Goverdhan, Michelle

Grunin, Shira Hagbi-Levi, Peter Campochiaro, Nicholas Katsanis, Frank G. Holz, Frédéric Blond, Hélène Blanché, Jean-François Deleuze, Robert P. Igo Jr, Barbara Truitt, Neal S. Peachey, Stacy M. Meuer, Chelsea E. Myers, Emily L. Moore, Ronald Klein, Michael A. Hauser, Eric A. Postel, Monique D. Courtenay, Stephen G. Schwartz, Jaclyn L. Kovach, William K. Scott, Gerald Liew, Ava G. Tan, Bamini Gopinath, John C. Merriam, R. Theodore Smith, Jane C. Khan, Humma Shahid, Anthony T. Moore, J. Allie McGrath, Reneé Laux, Milam A. Brantley Jr, Anita Agarwal, Lebriz Ersoy, Albert Caramoy, Thomas Langmann, Nicole T. M. Saksens, Eiko K. de Jong, Carel B. Hoyng, Melinda S. Cain, Andrea J. Richardson, Tammy M. Martin, John Blangero, Daniel E. Weeks, Bal Dhillon, Cornelia M. van Duijn, Kimberly F. Doheny, Jane Romm, Caroline C. W. Klaver, Caroline Hayward, Michael B. Gorin, Michael L. Klein, Paul N. Baird, Anneke I. den Hollander, Sascha Fauser, John R. W. Yates, Rando Allikmets, Jie Jin Wang, Debra A. Schaumberg, Barbara E. K. Klein, Stephanie A. Hagstrom, Itay Chowers, Andrew J. Lotery, Thierry Léveillard, Kang Zhang, Murray H. Brilliant, Alex W. Hewitt, Anand Swaroop, Emily Y. Chew, Margaret A. Pericak-Vance, Margaret DeAngelis, Dwight Stambolian, Jonathan L. Haines, Sudha K. Iyengar, Bernhard H. F. Weber, Gonçalo R. Abecasis, and Iris M. Heid. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics*, 48(2):134–143, February 2016.

[86] João M. Furtado, Justine R. Smith, Rubens Belfort, Devin Gattey, and Kevin L. Winthrop. Toxoplasmosis: A global threat. *Journal of Global Infectious Diseases*, 3(3):281–284, July 2011.

[87] A. M. García, T. Fletcher, F. G. Benavides, and E. Orts. Parental agricultural work and selected congenital malformations. *American Journal of Epidemiology*, 149(1):64–74, January 1999.

[88] LA Garcia-Cortes and D Sorensen. "Alternative implementations of Monte Carlo EM algorithms for likelihood inferences". *Genetics Selection Evolution: GSE*, 33(4):443–452, 2001.

[89] Vincent F Garry, Mary E Harkins, Leanna L Erickson, Leslie K Long-Simpson, Seth E Holland, and Barbara L Burroughs. Birth defects, season of conception, and sex of children born to pesticide applicators living in the Red River Valley of Minnesota, USA. *Environmental Health Perspectives*, 110(Suppl 3):441–449, June 2002.

[90] Paul A. Gastañaduy, Aron J. Hall, Aaron T. Curns, Umesh D. Parashar, and Benjamin A. Lopman. Burden of norovirus gastroenteritis in the ambulatory setting–United States, 2001-2009. *The Journal of Infectious Diseases*, 207(7):1058–1065, April 2013.

[91] William C. Gause, Thomas A. Wynn, and Judith E. Allen. Type 2 immunity and wound healing: Evolutionary refinement of adaptive immunity by helminths. *Nature Reviews Immunology*, 13(8):607–614, August 2013.

[92] Tian Ge, Chia-Yen Chen, Benjamin M. Neale, Mert R. Sabuncu, and Jordan W. Smoller. Phenome-wide Heritability Analysis of the UK Biobank. *bioRxiv*, page 070177, August 2016.

[93] A. Gelman and D.B. Rubin. Inference from Iterative Simulation using Multiple Sequences. *Statistical Science*, 7:457–511, 1992.

[94] Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, September 2006.

[95] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, Boca Raton, 3 edition edition, November 2013.

[96] US Census Bureau Geography. 2010 Census Urban Area Facts. `https://www.census.gov/geo/reference/ua/uafacts.html`.

[97] Noreen Goldman German Rodriguez. An Assessment of Estimation Procedures for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(1):73–89, 1995.

[98] R. Gijsen, N. Hoeymans, F. G. Schellevis, D. Ruwaard, W. A. Satariano, and G. A. van den Bos. Causes and consequences of comorbidity: A review. *Journal of Clinical Epidemiology*, 54(7):661–674, July 2001.

[99] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685 –8690, May 2007.

[100] Romina S. Goldszmid, Amiran Dzutsev, and Giorgio Trinchieri. Host immune response to infection and cancer: Unexpected commonalities. *Cell Host & Microbe*, 15(3):295–305, March 2014.

[101] Padhraig Gormley, Verneri Anttila, Bendik S Winsvold, Priit Palta, Tonu Esko, Tune H Pers, Kai-How Farh, Ester Cuenca-Leon, Mikko Muona, Nicholas A Furlotte, Tobias Kurth, Andres Ingason, George McMahon, Lannie Ligthart, Gisela M Terwindt, Mikko Kallela, Tobias M Freilinger, Caroline Ran, Scott G Gordon, Anine H Stam, Stacy Steinberg, Guntram Borck, Markku Koiranen, Lydia Quaye, Hieab H H Adams, Terho Lehtimaki, Antti-Pekka Sarin, Juho Wedenoja, David A Hinds, Julie E Buring, Markus Schurks, Paul M Ridker, Maria Gudlaug Hrafnsdottir, Hreinn Stefansson, Susan M Ring, Jouke-Jan Hottenga, Brenda W J H Penninx, Markus Farkkila, Ville Artto, Mari Kaunisto, Salli Vepsalainen, Rainer Malik, Andrew C Heath, Pamela A F Madden, Nicholas G Martin, Grant W Montgomery, Mitja I Kurki, Mart Kals, Reedik Magi, Kalle Parn, Eija Hamalainen, Hailiang Huang, Andrea E Byrnes, Lude Franke, Jie Huang, Evie Stergiakouli, Phil H Lee, Cynthia Sandor, Caleb Webber, Zameel Cader, Bertram Muller-Myhsok, Stefan Schreiber, Thomas Meitinger, Johan G Eriksson, Veikko Salomaa, Kauko Heikkila, Elizabeth Loehrer, Andre G Uitterlinden, Albert

Hofman, Cornelia M van Duijn, Lynn Cherkas, Linda M Pedersen, Audun Stubhaug, Christopher S Nielsen, Minna Mannikko, Evelin Mihailov, Lili Milani, Hartmut Gobel, Ann-Louise Esserlind, Anne Francke Christensen, Thomas Folkmann Hansen, Thomas Werge, International Headache Genetics Consortium, Jaakko Kaprio, Arpo J Aromaa, Olli Raitakari, M Arfan Ikram, Tim Spector, Marjo-Riitta Jarvelin, Andres Metspalu, Christian Kubisch, David P Strachan, Michel D Ferrari, Andrea C Belin, Martin Dichgans, Maija Wessman, Arn M J M van den Maagdenberg, John-Anker Zwart, Dorret I Boomsma, George Davey Smith, Kari Stefansson, Nicholas Eriksson, Mark J Daly, Benjamin M Neale, Jes Olesen, Daniel I Chasman, Dale R Nyholt, and Aarno Palotie. Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nat Genet*, advance online publication, June 2016.

[102] I. H. Grant, J. W. Gold, M. Rosenblum, D. Niedzwiecki, and D. Armstrong. Toxoplasma gondii serology in HIV-infected patients: The development of central nervous system toxoplasmosis in AIDS. *AIDS (London, England)*, 4(6):519–521, June 1990.

[103] Anna Grauers, Iffat Rahman, and Paul Gerdhem. Heritability of scoliosis. *European Spine Journal: Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, 21(6):1069–1074, June 2012.

[104] Casey S. Greene, Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, Boris M. Hartmann, Elena Zaslavsky, Stuart C. Sealfon, Daniel I. Chasman, Garret A. FitzGerald, Kara Dolinski, Tilo Grosser, and Olga G. Troyanskaya. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569–576, June 2015.

[105] Rebecca Grzadzinski, Marisela Huerta, and Catherine Lord. DSM-5 and autism spectrum disorders (ASDs): An opportunity for identifying ASD subtypes. *Molecular Autism*, 4:12, May 2013.

[106] Miriam L. Haaksma, Lara R. Vilela, Alessandra Marengoni, Amaia Calderón-Larrañaga, Jeannie-Marie S. Leoutsakos, Marcel G. M. Olde Rikkert, and René J. F. Melis. Comorbidity and progression of late onset Alzheimer's disease: A systematic review. *PloS One*, 12(5):e0177044, 2017.

[107] Jarrod D. Hadfield. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, 33:1–22, February 2010.

[108] J.D. Hadfield. *"MCMCglmm Course Notes"*. 2015.

[109] S. P. Hagenaars, S. E. Harris, G. Davies, W. D. Hill, D. C. M. Liewald, S. J. Ritchie, R. E. Marioni, C. Fawns-Ritchie, B. Cullen, R. Malik, B. B. Worrall, C. L. M. Sudlow, J. M. Wardlaw, J. Gallacher, J. Pell, A. M. McIntosh, D. J. Smith, C. R. Gale, and I. J. Deary. Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. *Molecular Psychiatry*, January 2016.

[110] Randi J. Hagerman and Paul Hagerman. Fragile X-associated tremor/ataxia syndrome - features, mechanisms and management. *Nature Reviews. Neurology*, 12(7):403–412, July 2016.

[111] Philip M. Hansbro, Gerard E. Kaiko, and Paul S. Foster. Cytokine/anti-cytokine therapy - novel treatments for asthma? *British Journal of Pharmacology*, 163(1):81–95, May 2011.

[112] U.S.A. Health Resources and Services Administration. AHRF: Area Health Resources Files. `http://www.arf.hrsa.gov/`.

[113] Andrew C. Heath, Dale R. Nyholt, Nathan A. Gillespie, and Nicholas G. Martin. Genetic Basis of Male Pattern Baldness. *Journal of Investigative Dermatology*, 121(6):1561–1564, December 2003.

[114] Donald Hedeker and Robert D. Gibbons. *Longitudinal Data Analysis*. Wiley-Interscience, Hoboken, N.J, 1 edition edition, April 2006.

[115] DR Hedeker and Robert D. Gibbons. SuperMix. `http://www.ssicentral.com/supermix/`.

[116] P. Heidelberger and PD. Welch. Simulation run length control in the presence of an initial transient. *Opns Res.*, 31:1109–44, 1983.

[117] Fred J. Hellinger and William E. Encinosa. The cost and incidence of prescribing errors among privately insured HIV patients. *PharmacoEconomics*, 28(1):23–34, 2010.

[118] M. L. Herdt-Losavio, S. Lin, B. R. Chapman, M. Hooiveld, A. Olshan, X. Liu, R. D. DePersis, J. Zhu, and C. M. Druschel. Maternal occupation and the risk of birth defects: An overview from the National Birth Defects Prevention Study. *Occupational and Environmental Medicine*, 67(1):58–66, January 2010.

[119] César A Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A Christakis. A dynamic network approach for the study of human phenotypes. *PLoS computational biology*, 5(4):e1000353, April 2009.

[120] Kate Hoffman, Amy E. Kalkbrenner, Veronica M. Vieira, and Julie L. Daniels. The spatial distribution of known predictors of autism spectrum disorders impacts geographic variability in prevalence in central North Carolina. *Environmental Health: A Global Access Science Source*, 11:80, October 2012.

[121] Karolina Holm, Johan Staaf, Martin Lauss, Mattias Aine, David Lindgren, Pär-Ola Bendahl, Johan Vallon-Christersson, Rosa Bjork Barkardottir, Mattias Höglund, Åke Borg, Göran Jönsson, and Markus Ringnér. An integrated genomics analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to chromatin states in normal mammary cells. *Breast Cancer Research : BCR*, 18, 2016.

[122] E. B. Hook. Cardiovascular birth defects and prenatal exposure to female sex hormones: A reevaluation of data reanalysis from a large prospective study. *Teratology*, 46(3):261–266, September 1992.

[123] E. B. Hook. Cardiovascular birth defects and prenatal exposure to female sex hormones: A reevaluation of data reanalysis from a large prospective study. *Teratology*, 49(3):162–166, March 1994.

[124] Richard S. Hotchkiss, Craig M. Coopersmith, Jonathan E. McDunn, and Thomas A. Ferguson. The sepsis seesaw: Tilting toward immunosuppression. *Nature Medicine*, 15(5):496–497, May 2009.

[125] Richard S. Hotchkiss, Guillaume Monneret, and Didier Payen. Sepsis-induced immunosuppression: From cellular dysfunctions to immunotherapy. *Nature Reviews Immunology*, 13(12):862–874, December 2013.

[126] Andrea Hotop, Harald Hlobil, and Uwe Gross. Efficacy of rapid treatment initiation following primary Toxoplasma gondii infection during pregnancy. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 54(11):1545–1552, June 2012.

[127] The White House. White House Precision Medicine Initiative. `https://obamawhitehouse.archives.gov/node/333101`.

[128] George Hripcsak and Adam S. Rothschild. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 12(3):296–298, 2005.

[129] Yueh-Han Hsu, Jur-Shan Cheng, Wen-Chen Ouyang, Chen-Li Lin, Chi-Ting Huang, and Chih-Cheng Hsu. Lower Incidence of End-Stage Renal Disease but Suboptimal Pre-Dialysis Renal Care in Schizophrenia: A 14-Year Nationwide Cohort Study. *PLoS ONE*, 10(10), October 2015.

[130] Jessica Xin Hu, Cecilia Engel Thomas, and Søren Brunak. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 17(10):615–629, October 2016.

[131] Hsuan-Hao Huang, Wei-Liang Shih, Yi-Hsiu Li, Chih-Feng Wu, Pei-Jer Chen, Chih-Lin Lin, Chun-Jen Liu, Yun-Fan Liaw, Shi-Ming Lin, Shou-Dong Lee, and Ming-Whei Yu. Hepatitis B viraemia: Its heritability and association with common genetic variation in the interferon $\gamma$ signalling pathway. *Gut*, 60(1):99–107, January 2011.

[132] Marc P. Hübner, Laura E. Layland, and Achim Hoerauf. Helminths and their implication in sepsis - a new branch of their immunomodulatory behaviour? *Pathogens and Disease*, 69(2):127–141, November 2013.

[133] David J. Hunter, Marlies de Lange, Harold Snieder, Alex J. MacGregor, R. Swaminathan, Rajesh V. Thakker, and Tim D. Spector. Genetic contribution to renal function and electrolyte balance: A twin study. *Clinical Science (London, England: 1979)*, 103(3):259–265, September 2002.

[134] Robin A. Hurley and Katherine H. Taber. Latent Toxoplasmosis gondii: Emerging evidence for influences on neuropsychiatric disorders. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 24(4):376–383, 2012.

[135] Kristina Ibáñez, César Boullosa, Rafael Tabarés-Seisdedos, Anaïs Baudot, and Alfonso Valencia. Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses. *PLoS genetics*, 10(2):e1004173, February 2014.

[136] Truls Ingebrigtsen, Simon F. Thomsen, Jørgen Vestbo, Sophie van der Sluis, Kirsten O. Kyvik, Edwin K. Silverman, Magnus Svartengren, and Vibeke Backer. Genetic influences on Chronic Obstructive Pulmonary Disease - a twin study. *Respiratory Medicine*, 104(12):1890–1895, December 2010.

[137] Johanne Telnes Instanes, Kari Klungsøyr, Anne Halmøy, Ole Bernt Fasmer, and Jan Haavik. Adult ADHD and Comorbid Somatic Disease: A Systematic Literature Review. *Journal of Attention Disorders*, September 2016.

[138] Ivan Iossifov, Michael Ronemus, Dan Levy, Zihua Wang, Inessa Hakker, Julie Rosenbaum, Boris Yamrom, Yoon-Ha Lee, Giuseppe Narzisi, Anthony Leotta, Jude Kendall, Ewa Grabowska, Beicong Ma, Steven Marks, Linda Rodgers, Asya Stepansky, Jennifer Troge, Peter Andrews, Mitchell Bekritsky, Kith Pradhan, Elena Ghiban, Melissa Kramer, Jennifer Parla, Ryan Demeter, Lucinda L. Fulton, Robert S. Fulton, Vincent J. Magrini, Kenny Ye, Jennifer C. Darnell, Robert B. Darnell, Elaine R. Mardis, Richard K. Wilson, Michael C. Schatz, W. Richard McCombie, and Michael Wigler. De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2):285–299, April 2012.

[139] A. Irgens, K. Krüger, A. H. Skorve, and L. M. Irgens. Birth defects and paternal occupational exposure. Hypotheses tested in a record linkage based dataset. *Acta Obstetricia Et Gynecologica Scandinavica*, 79(6):465–470, June 2000.

[140] Ida Jäderberg, Simon F. Thomsen, Kirsten O. Kyvik, Axel Skytthe, and Vibeke Backer. Atopic diseases in twins born after assisted reproduction. *Paediatric and Perinatal Epidemiology*, 26(2):140–145, March 2012.

[141] Anders Boeck Jensen, Pope L. Moseley, Tudor I. Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5:4022, June 2014.

[142] Jun-Tao Jiang, Wen-Lan Sun, Yi-Feng Jing, Shi-Bo Liu, Zheng Ma, Yan Hong, Long Ma, Chao Qin, Qiang Liu, Harrison J. Stratton, and Shu-Jie Xia. Prenatal exposure to

di-n-butyl phthalate induces anorectal malformations in male rat offspring. *Toxicology*, 290(2-3):322–326, December 2011.

[143] T. M. M. Joergensen, K. Christensen, J. S. Lindholt, L. A. Larsen, A. Green, and K. Houlind. High Heritability of Liability to Abdominal Aortic Aneurysms: A Population Based Twin Study. *Journal of Vascular Surgery*, 64(2):537, August 2016.

[144] J. L. Jones, D. L. Hanson, S. Y. Chu, C. A. Ciesielski, J. E. Kaplan, J. W. Ward, and T. R. Navin. Toxoplasmic encephalitis in HIV-infected persons: Risk factors and trends. The Adult/Adolescent Spectrum of Disease Group. *AIDS (London, England)*, 10(12):1393–1399, October 1996.

[145] Jeffrey L. Jones and Gary N. Holland. Annual burden of ocular toxoplasmosis in the US. *The American Journal of Tropical Medicine and Hygiene*, 82(3):464–465, March 2010.

[146] Jeffrey L. Jones, Deanna Kruszon-Moran, Hilda N. Rivera, Courtney Price, and Patricia P. Wilkins. Toxoplasma gondii seroprevalence in the United States 2009-2010 and comparison with the past two decades. *The American Journal of Tropical Medicine and Hygiene*, 90(6):1135–1139, June 2014.

[147] A. Kagan, B. R. Harris, W. Winkelstein, K. G. Johnson, H. Kato, S. L. Syme, G. G. Rhoads, M. L. Gay, M. Z. Nichaman, H. B. Hamilton, and J. Tillotson. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: Demographic, physical, dietary and biochemical characteristics. *Journal of Chronic Diseases*, 27(7-8):345–364, September 1974.

[148] Kenneth S. Kendler. The nature of psychiatric disorders. *World psychiatry: official journal of the World Psychiatric Association (WPA)*, 15(1):5–12, February 2016.

[149] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182, March 2003.

[150] Marissa D. King and Peter S. Bearman. Socioeconomic Status and the Increased Prevalence of Autism in California. *American Sociological Review*, 76(2):320–346, April 2011.

[151] M. J. Kjeldsen, K. O. Kyvik, K. Christensen, and M. L. Friis. Genetic and environmental factors in epilepsy: A population-based study of 11900 Danish twin pairs. *Epilepsy Research*, 44(2-3):167–178, May 2001.

[152] Ryan J. Koene, Anna E. Prizment, Anne Blaes, and Suma H. Konety. Shared Risk Factors in Cardiovascular Disease and Cancer. *Circulation*, 133(11):1104–1114, March 2016.

[153] Isaac S Kohane. Deeper, longer phenotyping to accelerate the discovery of the genetic architectures of diseases. *Genome Biology*, 15(5):115, 2014.

[154] I. R. Korsgaard, M. S. Lund, D. Sorensen, D. Gianola, P. Madsen, and J. Jensen. Multivariate Bayesian analysis of Gaussian, right censored Gaussian, ordered categorical and binary traits using Gibbs sampling. *Genetics Selection Evolution : GSE*, 35(2):159–183, 2003.

[155] Rebecca L. Koscik, Philip M. Farrell, Michael R. Kosorok, Kathleen M. Zaremba, Anita Laxova, Hui-Chuan Lai, Jeff A. Douglas, Michael J. Rock, and Mark L. Splaingard. Cognitive function of children with cystic fibrosis: Deleterious effect of early malnutrition. *Pediatrics*, 113(6):1549–1558, June 2004.

[156] Rose M. Kreider and Daphne A. Lofquist. *P20-572: Adopted Children and Stepchildren: 2010*. Washington, DC: U.S. Census Bureau, 2014.

[157] Paulette A Krishack, Kanix Wang, Andrey Rzhetsky, Julian Solway, Anne I Sperling, and Philip A. Verhoef. Pre-existing type 2 immune activation protects against the development of sepsis. *Journal of Allergy and Clinical Immunology*, 2017.

[158] Eswar Krishnan, Christina N. Lessov-Schlaggar, Ruth E. Krasnow, and Gary E. Swan. Nature Versus Nurture in Gout: A Twin Study. *The American Journal of Medicine*, 125(5):499–504, May 2012.

[159] J. Laherrère and D. Sornette. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B - Condensed Matter and Complex Systems*, 2(4):525–539, April 1998.

[160] Alison E. Lane, Robyn L. Young, Amy E. Z. Baker, and Manya T. Angley. Sensory Processing Subtypes in Autism: Association with Adaptive Behavior. *Journal of Autism and Developmental Disorders*, 40(1):112–122, January 2010.

[161] Heidi Jeanet Larsson, William W. Eaton, Kreesten Meldgaard Madsen, Mogens Vestergaard, Anne Vingaard Olesen, Esben Agerbo, Diana Schendel, Poul Thorsen, and Preben Bo Mortensen. Risk factors for autism: Perinatal factors, parental psychiatric history, and socioeconomic status. *American Journal of Epidemiology*, 161(10):916–925; discussion 926–928, May 2005.

[162] P. Latkany. 5 - Ocular Disease Due to Toxoplasma gondii. In *Toxoplasma Gondii*, pages 101–131. Academic Press, London, 2007.

[163] Joel Lavinsky, Amanda L. Crow, Calvin Pan, Juemei Wang, Ksenia A. Aaron, Maria K. Ho, Qingzhong Li, Pehzman Salehide, Anthony Myint, Maya Monges-Hernadez, Eleazar Eskin, Hooman Allayee, Aldons J. Lusis, and Rick A. Friedman. Genome-wide association study identifies nox3 as a critical gene for susceptibility to noise-induced hearing loss. *PLoS genetics*, 11(4):e1005094, April 2015.

[164] J. S. Lawrence, C. L. Martins, and G. L. Drake. A family survey of lupus erythematosus. 1. Heritability. *The Journal of Rheumatology*, 14(5):913–921, October 1987.

[165] Christina C. Lawson, Teresa M. Schnorr, Elizabeth A. Whelan, James A. Deddens, David A. Dankovic, Laurie A. Piacitelli, Marie H. Sweeney, and L. Barbara Connally. Paternal occupational exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin and birth outcomes of offspring: Birth weight, preterm delivery, and birth defects. *Environmental Health Perspectives*, 112(14):1403–1408, October 2004.

[166] Maud Lélu, Isabelle Villena, Marie-Laure Dardé, Dominique Aubert, Régine Geers, Emilie Dupuis, Francine Marnef, Marie-Lazarine Poulle, Cécile Gotteland, Aurélien Dumètre, and Emmanuelle Gilot-Fromont. Quantitative estimation of the viability of Toxoplasma gondii oocysts in soil. *Applied and Environmental Microbiology*, 78(15):5127–5132, August 2012.

[167] T. A. Lewandowski, S. M. Bartell, J. W. Yager, and L. Levin. An evaluation of surrogate chemical exposure measures and autism prevalence in Texas. *Journal of Toxicology and Environmental Health. Part A*, 72(24):1592–1603, 2009.

[168] Chun-Hui Liao, Chen-Shu Chang, Wan-Ching Wei, Shih-Ni Chang, Chien-Chang Liao, Hsien-Yuan Lane, and Fung-Chang Sung. Schizophrenia patients at higher risk of diabetes, hypertension and hyperlipidemia: A population-based study. *Schizophrenia Research*, 126(1-3):110–116, March 2011.

[169] P Lichtenstein, BH Yip, C Björk, Y Pawitan, TD Cannon, PF Sullivan, and CM. Hultman. "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: A population-based study.". *Lancet*, 373(9659):234–9, January 2009.

[170] Shao Lin, Michele L. Herdt-Losavio, Bonnie R. Chapman, Jean-Pierre Munsie, Andrew F. Olshan, Charlotte M. Druschel, and National Birth Defects Prevention Study. Maternal occupation and the risk of major birth defects: A follow-up analysis from the National Birth Defects Prevention Study. *International Journal of Hygiene and Environmental Health*, 216(3):317–323, June 2013.

[171] Chunyu Liu, Josée Dupuis, Martin G. Larson, L. Adrienne Cupples, Jose M. Ordovas, Ramachandran S. Vasan, James B. Meigs, Paul F. Jacques, and Daniel Levy. Revisiting heritability accounting for shared environmental effects and maternal inheritance. *Human Genetics*, 134(2):169–179, February 2015.

[172] Jimmy Z Liu, Mohamed A Almarri, Daniel J Gaffney, George F Mells, Luke Jostins, Heather J Cordell, Samantha J Ducker, Darren B Day, Michael A Heneghan, James M. Neuberger, Peter T Donaldson, Andrew J Bathgate, Andrew Burroughs, Mervyn H Davies, David E Jones, Graeme J Alexander, Jeffrey C Barrett, Richard N Sandford, and Carl A Anderson. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nature genetics*, 44(10):1137–1141, October 2012.

[173] Jimmy Z Liu, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, Tejas Shah, Shifteh Abedian, Jae Hee Cheon, Judy Cho, Naser E Dayani, Lude Franke, Yuta Fuyuno, Ailsa Hart, Ramesh C Juyal, Garima Juyal, Won Ho Kim, Andrew P Morris, Hossein

Poustchi, William G Newman, Vandana Midha, Timothy R Orchard, Homayon Vahedi, Ajit Sood, Joseph Y Sung, Reza Malekzadeh, Harm-Jan Westra, Keiko Yamazaki, Suk-Kyun Yang, Jeffrey C Barrett, Behrooz Z Alizadeh, Miles Parkes, Thelma BK, Mark J Daly, Michiaki Kubo, Carl A Anderson, and Rinse K Weersma. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979–986, September 2015.

[174] XF Liu and MJ Daniels. "A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization". *Journal of Computational and Graphical Statistics*, 15(4):897–914, 2006.

[175] P-R Loh, G Bhatia, A Gusev, and et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nature genetics*, 47(12):1385–1392, 2015.

[176] N. Lomtatidze, G. Dumbadze, M. Chkhaidze, and R. Khakhnelidze. IMPACT OF SOME ENVIRONMENTAL FACTORS ON HUMAN HEALTH. *Georgian Medical News*, (258):64–67, September 2016.

[177] Rohit Loomba, Nicholas Schork, Chi-Hua Chen, Ricki Bettencourt, Ana Bhatt, Brandon Ang, Phirum Nguyen, Carolyn Hernandez, Lisa Richards, Joanie Salotti, Steven Lin, Ekihiro Seki, Karen E. Nelson, Claude B. Sirlin, and David Brenner. Heritability of Hepatic Fibrosis and Steatosis Based on a Prospective Twin Study. *Gastroenterology*, 149(7):1784–1793, December 2015.

[178] Yang Luo, Katrina M. de Lange, Luke Jostins, Loukas Moutsianas, Joshua Randall, Nicholas A. Kennedy, Christopher A. Lamb, Shane McCarthy, Tariq Ahmad, Cathryn Edwards, Eva Goncalves Serra, Ailsa Hart, Chris Hawkey, John C. Mansfield, Craig Mowat, William G. Newman, Sam Nichols, Martin Pollard, Jack Satsangi, Alison Simmons, Mark Tremelling, Holm Uhlig, David C. Wilson, James C. Lee, Natalie J. Prescott, Charlie W. Lees, Christopher G. Mathew, Miles Parkes, Jeffrey C. Barrett, and Carl A. Anderson. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nature Genetics*, 49(2):186–192, February 2017.

[179] Joseph Lykins, Kanix Wang, Kelsey Wheeler, Fatima Clouser, Ashtyn Dixon, Kamal El Bissati, Ying Zhou, Christopher Lyttle, Andrey Rzhetsky, and Rima McLeod. Understanding Toxoplasmosis in the United States Through "Large Data" Analyses. *Clinical Infectious Diseases*, 63(4):468–475, August 2016.

[180] John Lynch and George Davey Smith. A Life Course Approach to Chronic Disease Epidemiology. *Annual Review of Public Health*, 26(1):1–35, 2005.

[181] Michael Lynch and Bruce Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, Mass, 1 edition edition, January 1998.

[182] Jaclyn R. MacFarlane and Tomoe Kanaya. What Does it Mean to be Autistic? Interstate Variation in Special Education Criteria for Autism Services. *Journal of Child and Family Studies*, 18(6):662, December 2009.

[183] A. J. MacGregor, H. Snieder, A. S. Rigby, M. Koskenvuo, J. Kaprio, K. Aho, and A. J. Silman. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis and Rheumatism*, 43(1):30–37, January 2000.

[184] Mario Maj. "Psychiatric comorbidity": An artefact of current diagnostic systems? *The British Journal of Psychiatry: The Journal of Mental Science*, 186:182–184, March 2005.

[185] Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Stephan Ripke, Naomi R. Wray, Cathryn M. Lewis, Steven P. Hamilton, Myrna M. Weissman, Gerome Breen, Enda M. Byrne, Douglas H. R. Blackwood, Dorret I. Boomsma, Sven Cichon, Andrew C. Heath, Florian Holsboer, Susanne Lucae, Pamela A. F. Madden, Nicholas G. Martin, Peter McGuffin, Pierandrea Muglia, Markus M. Noethen, Brenda P. Penninx, Michele L. Pergadia, James B. Potash, Marcella Rietschel, Danyu Lin, Bertram Müller-Myhsok, Jianxin Shi, Stacy Steinberg, Hans J. Grabe, Paul Lichtenstein, Patrik Magnusson, Roy H. Perlis, Martin Preisig, Jordan W. Smoller, Kari Stefansson, Rudolf Uher, Zoltan Kutalik, Katherine E. Tansey, Alexander Teumer, Alexander Viktorin, Michael R. Barnes, Thomas Bettecken, Elisabeth B. Binder, René Breuer, Victor M. Castro, Susanne E. Churchill, William H. Coryell, Nick Craddock, Ian W. Craig, Darina Czamara, Eco J. De Geus, Franziska Degenhardt, Anne E. Farmer, Maurizio Fava, Josef Frank, Vivian S. Gainer, Patience J. Gallagher, Scott D. Gordon, Sergey Goryachev, Magdalena Gross, Michel Guipponi, Anjali K. Henders, Stefan Herms, Ian B. Hickie, Susanne Hoefels, Witte Hoogendijk, Jouke Jan Hottenga, Dan V. Iosifescu, Marcus Ising, Ian Jones, Lisa Jones, Tzeng Jung-Ying, James A. Knowles, Isaac S. Kohane, Martin A. Kohli, Ania Korszun, Mikael Landen, William B. Lawson, Glyn Lewis, Donald Macintyre, Wolfgang Maier, Manuel Mattheisen, Patrick J. McGrath, Andrew McIntosh, Alan McLean, Christel M. Middeldorp, Lefkos Middleton, Grant M. Montgomery, Shawn N. Murphy, Matthias Nauck, Willem A. Nolen, Dale R. Nyholt, Michael O'Donovan, Högni Oskarsson, Nancy Pedersen, William A. Scheftner, Andrea Schulz, Thomas G. Schulze, Stanley I. Shyn, Engilbert Sigurdsson, Susan L. Slager, Johannes H. Smit, Hreinn Stefansson, Michael Steffens, Thorgeir Thorgeirsson, Federica Tozzi, Jens Treutlein, Manfred Uhr, Edwin J. C. G. van den Oord, Gerard Van Grootheest, Henry Völzke, Jeffrey B. Weilburg, Gonneke Willemsen, Frans G. Zitman, Benjamin Neale, Mark Daly, Douglas F. Levinson, and Patrick F. Sullivan. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, 18(4):497–511, April 2013.

[186] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[187] C. Mathieu, M. Waer, J. Laureys, O. Rutgeerts, and R. Bouillon. Prevention of autoimmune diabetes in NOD mice by 1,25 dihydroxyvitamin D3. *Diabetologia*, 37(6):552–558, June 1994.

[188] Sarah N. Mattson and Edward P. Riley. A Review of the Neurobehavioral Deficits in Children with Fetal Alcohol Syndrome or Prenatal Exposure to Alcohol. *Alcoholism: Clinical and Experimental Research*, 22(2):279–294, April 1998.

[189] Rima McLeod, Joseph Lykins, A. Gwendolyn Noble, Peter Rabiah, Charles N. Swisher, Peter T. Heydemann, David McLone, David Frim, Shawn Withers, Fatima Clouser, and Kenneth Boyer. Management of Congenital Toxoplasmosis. *Current Pediatrics Reports*, 2(3):166–194, September 2014.

[190] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science (New York, N.Y.)*, 347(6224):1257601, February 2015.

[191] D. Mickley, D. Greenfeld, D. M. Quinlan, P. Roloff, and F. Zwas. Abnormal liver enzymes in outpatients with eating disorders. *The International Journal of Eating Disorders*, 20(3):325–329, November 1996.

[192] Ozlem Miman, Ozge Yilmaz Kusbeci, Orhan Cem Aktepe, and Zafer Cetinkaya. The probable relation between Toxoplasma gondii and Parkinson's disease. *Neuroscience Letters*, 475(3):129–131, May 2010.

[193] Miriam F. Moffatt, Ivo G. Gut, Florence Demenais, David P. Strachan, Emmanuelle Bouzigon, Simon Heath, Erika von Mutius, Martin Farrall, Mark Lathrop, and William O.C.M. Cookson. A Large-Scale, Consortium-Based Genomewide Association Study of Asthma. *New England Journal of Medicine*, 363(13):1211–1221, September 2010.

[194] Sukhleen K. Momi, Lisa E. Wolber, Stella Maris Fabiane, Alex J. MacGregor, and Frances M. K. Williams. Genetic and Environmental Factors in Age-Related Hearing Impairment. *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies*, 18(4):383–392, August 2015.

[195] Amy Moore and Daniel A. Enquobahrie. Paternal occupational exposure to pesticides and risk of neuroblastoma among children: A meta-analysis. *Cancer causes & control: CCC*, 22(11):1529–1536, November 2011.

[196] MB Morrissey, AJ Wilson, JM Pemberton, and MM Ferguson. A framework for power and sensitivity analyses for quantitative genetic studies of natural populations, and case studies in Soay sheep (Ovis aries). *Journal of Evolutionary Biology*, 20(2):2309–2321, 2007.

[197] LA Mucci, JB Hjelmborg, JR Harris, and et al. Familial risk and heritability of cancer among twins in nordic countries. *JAMA*, 315(1):68–76, 2016.

[198] María Muñoz, Ricardo Pong-Wong, Oriol Canela-Xandri, Konrad Rawlik, Chris S. Haley, and Albert Tenesa. Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank. *Nature Genetics*, 48(9):980–983, September 2016.

[199] Nubia Muñoz, F. Xavier Bosch, Silvia de Sanjosé, Rolando Herrero, Xavier Castellsagué, Keerti V. Shah, Peter J.F. Snijders, and Chris J.L.M. Meijer. Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer. *New England Journal of Medicine*, 348(6):518–527, February 2003.

[200] Darren A. Natale, Cecilia N. Arighi, Winona C. Barker, Judith A. Blake, Carol J. Bult, Michael Caudy, Harold J. Drabkin, Peter D'Eustachio, Alexei V. Evsikov, Hongzhan Huang, Jules Nchoutmboube, Natalia V. Roberts, Barry Smith, Jian Zhang, and Cathy H. Wu. The Protein Ontology: A structured representation of protein forms and complexes. *Nucleic Acids Research*, 39(Database issue):D539–D545, January 2011.

[201] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC), 2011.

[202] Benjamin M. Neale, Sarah E. Medland, Stephan Ripke, Philip Asherson, Barbara Franke, Klaus-Peter Lesch, Stephen V. Faraone, Thuy Trang Nguyen, Helmut Schäfer, Peter Holmans, Mark Daly, Hans-Christoph Steinhausen, Christine Freitag, Andreas Reif, Tobias J. Renner, Marcel Romanos, Jasmin Romanos, Susanne Walitza, Andreas Warnke, Jobst Meyer, Haukur Palmason, Jan Buitelaar, Alejandro Arias Vasquez, Nanda Lambregts-Rommelse, Michael Gill, Richard J. L. Anney, Kate Langely, Michael O'Donovan, Nigel Williams, Michael Owen, Anita Thapar, Lindsey Kent, Joseph Sergeant, Herbert Roeyers, Eric Mick, Joseph Biederman, Alysa Doyle, Susan Smalley, Sandra Loo, Hakon Hakonarson, Josephine Elia, Alexandre Todorov, Ana Miranda, Fernando Mulas, Richard P. Ebstein, Aribert Rothenberger, Tobias Banaschewski, Robert D. Oades, Edmund Sonuga-Barke, James McGough, Laura Nisenbaum, Frank Middleton, Xiaolan Hu, Stan Nelson, and Psychiatric GWAS Consortium: ADHD Subgroup. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 49(9):884–897, September 2010.

[203] The Cancer Genome Atlas Research Network. Integrated Genomic Analyses of Ovarian Carcinoma. *Nature*, 474(7353):609–615, June 2011.

[204] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, September 2012.

[205] Mariana Neves. An analysis on the entity annotations in biological corpora. *F1000Research*, 3, April 2014.

[206] Molly A. Nikolas and S. Alexandra Burt. Genetic and environmental influences on ADHD symptom dimensions of inattention and hyperactivity: A meta-analysis. *Journal of Abnormal Psychology*, 119(1):1–17, 2010.

[207] Sharon-Lise T. Normand, Mary Beth Landrum, Edward Guadagnoli, John Z. Ayanian, Thomas J. Ryan, Paul D. Cleary, and Barbara J. McNeil. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54(4):387 – 398, 2001.

[208] T. Norström and O. J. Skog. Alcohol and mortality: Methodological and analytical issues in aggregate analyses. *Addiction (Abingdon, England)*, 96 Suppl 1:S5–17, February 2001.

[209] Shannon A. Novosad, Mathew R. P. Sapiano, Cheri Grigg, Jason Lake, Misha Robyn, Ghinwa Dumyati, Christina Felsen, Debra Blog, Elizabeth Dufort, Shelley Zansky, Kathryn Wiedeman, Lacey Avery, Raymund B. Dantes, John A. Jernigan, Shelley S. Magill, Anthony Fiore, and Lauren Epstein. Vital Signs: Epidemiology of Sepsis: Prevalence of Health Care Factors and Opportunities for Prevention. *MMWR. Morbidity and mortality weekly report*, 65(33):864–869, August 2016.

[210] Keisuke Oboki, Tatsukuni Ohno, Naoki Kajiwara, Ken Arae, Hideaki Morita, Akina Ishii, Aya Nambu, Takaya Abe, Hiroshi Kiyonari, Kenji Matsumoto, Katsuko Sudo, Ko Okumura, Hirohisa Saito, and Susumu Nakae. IL-33 is a crucial amplifier of innate rather than acquired immunity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(43):18581–18586, October 2010.

[211] Asmundur Oddsson, Patrick Sulem, Hannes Helgason, Vidar O. Edvardsson, Gudmar Thorleifsson, Gardar Sveinbjörnsson, Eik Haraldsdottir, Gudmundur I. Eyjolfsson, Olof Sigurdardottir, Isleifur Olafsson, Gisli Masson, Hilma Holm, Daniel F. Gudbjartsson, Unnur Thorsteinsdottir, Olafur S. Indridason, Runolfur Palsson, and Kari Stefansson. Common and rare variants associated with kidney stones and biochemical traits. *Nature Communications*, 6:7975, August 2015.

[212] Jelili Ojodu, Mary M. Hulihan, Shammara N. Pope, Althea M. Grant, and Centers for Disease Control and Prevention (CDC). Incidence of sickle cell trait–United States, 2010. *MMWR. Morbidity and mortality weekly report*, 63(49):1155–1158, December 2014.

[213] Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, Robert R. Graham, Arun Manoharan, Ward Ortmann, Tushar Bhangale, Joshua C. Denny, Robert J. Carroll, Anne E. Eyler, Jeffrey D. Greenberg, Joel M. Kremer, Dimitrios A. Pappas, Lei Jiang, Jian Yin, Lingying Ye, Ding-Feng Su, Jian Yang, Gang Xie, Ed Keystone, Harm-Jan Westra, Tõnu Esko, Andres Metspalu, Xuezhong Zhou, Namrata Gupta, Daniel Mirel, Eli A. Stahl, Dorothée Diogo, Jing Cui, Katherine Liao, Michael H. Guo, Keiko Myouzen, Takahisa Kawaguchi, Marieke J.H. Coenen, Piet L.C.M. van

Riel, Mart A.F.J. van de Laar, Henk-Jan Guchelaar, Tom W.J. Huizinga, Philippe Dieudé, Xavier Mariette, S. Louis Bridges, Alexandra Zhernakova, Rene E.M. Toes, Paul P. Tak, Corinne Miceli-Richard, So-Young Bang, Hye-Soon Lee, Javier Martin, Miguel A. Gonzalez-Gay, Luis Rodriguez-Rodriguez, Solbritt Rantapää-Dahlqvist, Lisbeth Ärlestig, Hyon K. Choi, Yoichiro Kamatani, Pilar Galan, Mark Lathrop, Steve Eyre, John Bowes, Anne Barton, Niek de Vries, Larry W. Moreland, Lindsey A. Criswell, Elizabeth W. Karlson, Atsuo Taniguchi, Ryo Yamada, Michiaki Kubo, Jun S. Liu, Sang-Cheol Bae, Jane Worthington, Leonid Padyukov, Lars Klareskog, Peter K. Gregersen, Soumya Raychaudhuri, Barbara E. Stranger, Philip L. De Jager, Lude Franke, Peter M. Visscher, Matthew A. Brown, Hisashi Yamanaka, Tsuneyo Mimori, Atsushi Takahashi, Huji Xu, Timothy W. Behrens, Katherine A. Siminovitch, Shigeki Momohara, Fumihiko Matsuda, Kazuhiko Yamamoto, and Robert M. Plenge. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, February 2014.

[214] Darren D. O'Rielly, Mohammed Uddin, and Proton Rahman. Ankylosing spondylitis: Beyond genome-wide association studies. *Current Opinion in Rheumatology*, 28(4):337–345, July 2016.

[215] Brian J. O'Roak, Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P. Coe, Roie Levy, Arthur Ko, Choli Lee, Joshua D. Smith, Emily H. Turner, Ian B. Stanaway, Benjamin Vernot, Maika Malig, Carl Baker, Beau Reilly, Joshua M. Akey, Elhanan Borenstein, Mark J. Rieder, Deborah A. Nickerson, Raphael Bernier, Jay Shendure, and Evan E. Eichler. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–250, April 2012.

[216] Kwame Owusu-Edusei, Harrell W. Chesson, and Thomas L. Gift. The economic burden of pediculosis pubis and scabies infections treated on an outpatient basis in the United States: Evidence from private insurance claims data, 2001-2005. *Sexually Transmitted Diseases*, 36(5):297–299, May 2009.

[217] Kwame Owusu-Edusei, Thomas L. Gift, and Harrell W. Chesson. Treatment cost of acute gonococcal infections: Estimates from employer-sponsored private insurance claims data in the United States, 2003-2007. *Sexually Transmitted Diseases*, 37(5):316–318, May 2010.

[218] H. M. Ozgen, J. W. Hop, J. J. Hox, F. A. Beemer, and H. van Engeland. Minor physical anomalies in autism: A meta-analysis. *Molecular Psychiatry*, 15(3):300–307, March 2010.

[219] Orazio Palumbo, Pietro Palumbo, Raffaella Stallone, Teresa Palladino, Leopoldo Zelante, and Massimo Carella. 8q12.1q12.3 de novo microdeletion involving the CHD7 gene in a patient without the major features of CHARGE syndrome: Case report and critical review of the literature. *Gene*, 513(1):209–213, January 2013.

[220] Bogdan Pasaniuc and Alkes L. Price. Dissecting the genetics of complex traits using summary association statistics. *bioRxiv*, page 072934, September 2016.

[221] S. R. Patel, E. K. Larkin, and S. Redline. Shared genetic basis for obstructive sleep apnea and adiposity measures. *International Journal of Obesity (2005)*, 32(5):795–800, May 2008.

[222] Ariana Peck and Elizabeth D. Mellins. Precarious balance: Th17 cells in host defense. *Infection and Immunity*, 78(1):32–38, January 2010.

[223] Judith Pinborough-Zimmerman, Deborah Bilder, Amanda Bakian, Robert Satterfield, Paul S. Carbone, Barry E. Nangle, Harper Randall, and William M. McMahon. Sociodemographic risk factors associated with autism spectrum disorders and intellectual disability. *Autism Research: Official Journal of the International Society for Autism Research*, 4(6):438–448, December 2011.

[224] Karly Pippitt, Marlana Li, and Holly E. Gurgle. Diabetes Mellitus: Screening and Diagnosis. *American Family Physician*, 93(2):103–109, January 2016.

[225] M Plummer, N Best, K Cowles, and K Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6:7–11, 2006.

[226] TJ Polderman, B Benyamin, CA de Leeuw, PF Sullivan, A van Bochoven, PM Visscher, and D Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, 47(7):702–9, 2015.

[227] Eleonora Porcu, Marco Medici, Giorgio Pistis, Claudia B. Volpato, Scott G. Wilson, Anne R. Cappola, Steffan D. Bos, Joris Deelen, Martin den Heijer, Rachel M. Freathy, Jari Lahti, Chunyu Liu, Lorna M. Lopez, Ilja M. Nolte, Jeffrey R. O'Connell, Toshiko Tanaka, Stella Trompet, Alice Arnold, Stefania Bandinelli, Marian Beekman, Stefan Böhringer, Suzanne J. Brown, Brendan M. Buckley, Clara Camaschella, Anton J. M. de Craen, Gail Davies, Marieke C. H. de Visser, Ian Ford, Tom Forsen, Timothy M. Frayling, Laura Fugazzola, Martin Gögele, Andrew T. Hattersley, Ad R. Hermus, Albert Hofman, Jeanine J. Houwing-Duistermaat, Richard A. Jensen, Eero Kajantie, Margreet Kloppenburg, Ee M. Lim, Corrado Masciullo, Stefano Mariotti, Cosetta Minelli, Braxton D. Mitchell, Ramaiah Nagaraja, Romana T. Netea-Maier, Aarno Palotie, Luca Persani, Maria G. Piras, Bruce M. Psaty, Katri Räikkönen, J. Brent Richards, Fernando Rivadeneira, Cinzia Sala, Mona M. Sabra, Naveed Sattar, Beverley M. Shields, Nicole Soranzo, John M. Starr, David J. Stott, Fred C. G. J. Sweep, Gianluca Usala, Melanie M. van der Klauw, Diana van Heemst, Alies van Mullem, Sita H.Vermeulen, W. Edward Visser, John P. Walsh, Rudi G. J. Westendorp, Elisabeth Widen, Guangju Zhai, Francesco Cucca, Ian J. Deary, Johan G. Eriksson, Luigi Ferrucci, Caroline S. Fox, J. Wouter Jukema, Lambertus A. Kiemeney, Peter P. Pramstaller, David Schlessinger, Alan R. Shuldiner, Eline P. Slagboom, André G. Uitterlinden, Bijay Vaidya, Theo J. Visser, Bruce H. R. Wolffenbuttel, Ingrid Meulenbelt, Jerome I. Rotter, Tim D. Spector, Andrew A. Hicks, Daniela Toniolo, Serena Sanna, Robin P. Peeters, and Silvia Naitza. A Meta-Analysis of Thyroid-Related Traits Reveals Novel Loci and Gender-Specific Differences in the Regulation of Thyroid Function. *PLoS Genetics*, 9(2), February 2013.

[228] S. B. Porter and M. A. Sande. Toxoplasmosis of the central nervous system in the acquired immunodeficiency syndrome. *The New England Journal of Medicine*, 327(23):1643–1648, December 1992.

[229] Michael E. Powers, Russell E. N. Becker, Anne Sailer, Jerrold R. Turner, and Juliane Bubeck Wardenburg. Synergistic Action of Staphylococcus aureus $\alpha$-Toxin on Platelets and Myeloid Lineage Cells Contributes to Lethal Sepsis. *Cell Host & Microbe*, 17(6):775–787, June 2015.

[230] Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics*, 43(10):977–983, September 2011.

[231] Andres J. Pumariega and Joseph D. LaBarbera. Eating attitudes and personality variables in a nonclinical sample. *International Journal of Eating Disorders*, 5(2):285–294, February 1986.

[232] S Rabe-Hesketh, A Skrondal, and A Pickles. GLLAMM: Generalized Linear Latent And Mixed Models. 2013.

[233] NANCY C. RAYMOND, PI-NIAN CHANG, SCOTT J. CROW, JAMES E. MITCHELL, BENITA S. DIEPERINK, MELISSA M. BECK, ROSS D. CROSBY, C. CARLYLE CLAWSON, and WARREN J. WARWICK. Eating disorders in patients with cystic fibrosis. *Journal of Adolescence*, 23(3):359–363, June 2000.

[234] P. J. Reitnauer, N. P. Callanan, R. A. Farber, and A. S. Aylsworth. Prenatal exposure to disulfiram implicated in the cause of malformations in discordant monozygotic twins. *Teratology*, 56(6):358–362, December 1997.

[235] R. Repetto and S. S. Baliga. Pesticides and the immune system: The public health risks. Executive summary. *Central European Journal of Public Health*, 4(4):263–265, December 1996.

[236] Elise B. Robinson, Beate St. Pourcain, Verneri Anttila, Jack A. Kosmicki, Brendan Bulik-Sullivan, Jakob Grove, Julian Maller, Kaitlin E. Samocha, Stephan J. Sanders, Stephan Ripke, Joanna Martin, Mads V. Hollegaard, Thomas Werge, David M. Hougaard, Benjamin M. Neale, David M. Evans, David Skuse, Preben Bo Mortensen, Anders D. Børglum, Angelica Ronald, George Davey Smith, and Mark J. Daly. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nature genetics*, 48(5):552–555, May 2016.

[237] Federico Rojo, Abel González-Pérez, Jessica Furriol, Ma Jesús Nicolau, Jaime Ferrer, Octavio Burgués, MohammadA Sabbaghi, Irene González-Navarrete, Ion Cristobal, Laia Serrano, Sandra Zazo, Juan Madoz, Sonia Servitja, Ignasi Tusquets, Joan Albanell, Ana Lluch, Ana Rovira, and Pilar Eroles. Non-canonical NF-$\kappa$B pathway activation predicts outcome in borderline oestrogen receptor positive breast carcinoma. *British Journal of Cancer*, 115(3):322–331, July 2016.

[238] T. Roos, J. Martius, U. Gross, and L. Schrod. Systematic serologic screening for toxoplasmosis in pregnancy. *Obstetrics and Gynecology*, 81(2):243–250, February 1993.

[239] Maroeska Rovers, Mark Haggard, Mary Gannon, Gesina Koeppen-Schomerus, and Robert Plomin. Heritability of Symptom Domains in Otitis Media: A Longitudinal Study of 1,373 Twin Pairs. *American Journal of Epidemiology*, 155(10):958–964, May 2002.

[240] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, October 2005.

[241] D.B. Rubin. Estimating the causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psych.*, 66:688–701, 1974.

[242] DonaldB. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188, 2001.

[243] Jeffrey D. Rudie, Leanna M. Hernandez, Jesse A. Brown, Devora Beck-Pancer, Natalie L. Colich, Philip Gorrindo, Paul M. Thompson, Daniel H. Geschwind, Susan Y. Bookheimer, Pat Levitt, and Mirella Dapretto. Autism-associated promoter variant in MET impacts functional and structural brain networks. *Neuron*, 75(5):904–915, September 2012.

[244] Andrey Rzhetsky, Steven C. Bagley, Kanix Wang, Christopher S. Lyttle, Edwin H. Cook Jr, Russ B. Altman, and Robert D. Gibbons. Environmental and State-Level Regulatory Factors Affect the Incidence of Autism and Intellectual Disability. *PLOS Computational Biology*, 10(3):e1003518, March 2014.

[245] O. Sadr Azodi, Å. Andrén-Sandberg, and H. Larsson. Genetic and environmental influences on the risk of acute appendicitis in twins. *British Journal of Surgery*, 96(11):1336–1340, November 2009.

[246] Ulrich Sagel, Alexander Krämer, and Rafael T. Mikolajczyk. "Blind periods" in screening for toxoplasmosis in pregnancy in Austria - a debate. *BMC infectious diseases*, 12:118, May 2012.

[247] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, 1987.

[248] Masakiyo Sakaguchi, Hiroyuki Sonegawa, Takamasa Nukui, Yoshihiko Sakaguchi, Masahiro Miyazaki, Masayoshi Namba, and Nam-ho Huh. Bifurcated converging pathways for high Ca2+- and TGF$\beta$-induced inhibition of growth of normal human keratinocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13921–13926, September 2005.

[249] M. Sallmén, M. L. Lindbohm, A. Anttila, H. Taskinen, and K. Hemminki. Paternal occupational lead exposure and congenital malformations. *Journal of Epidemiology and Community Health*, 46(5):519–522, October 1992.

[250] Francisco-Javier San-Andrés, Rafael Rubio, Jesús Castilla, Federico Pulido, Guillermo Palao, Inmaculada de Pedro, José-Ramón Costa, and Angel del Palacio. Incidence of acquired immunodeficiency syndrome-associated opportunistic diseases and the effect of treatment on a cohort of 1115 patients infected with human immunodeficiency virus, 1989-1997. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 36(9):1177–1185, May 2003.

[251] Stephan J. Sanders, Michael T. Murtha, Abha R. Gupta, John D. Murdoch, Melanie J. Raubeson, A. Jeremy Willsey, A. Gulhan Ercan-Sencicek, Nicholas M. DiLullo, Neelroop N. Parikshak, Jason L. Stein, Michael F. Walker, Gordon T. Ober, Nicole A. Teran, Youeun Song, Paul El-Fishawy, Ryan C. Murtha, Murim Choi, John D. Overton, Robert D. Bjornson, Nicholas J. Carriero, Kyle A. Meyer, Kaya Bilguvar, Shrikant M. Mane, Nenad Sestan, Richard P. Lifton, Murat Günel, Kathryn Roeder, Daniel H. Geschwind, Bernie Devlin, and Matthew W. State. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–241, April 2012.

[252] Paul G. Sanfilippo, Alex W. Hewitt, Chris J. Hammond, and David A. Mackey. The Heritability of Ocular Traits. *Survey of Ophthalmology*, 55(6):561–583, November 2010.

[253] Carolyn E. Sartor, Julia D. Grant, Michael T. Lynskey, Vivia V. McCutcheon, Mary Waldron, Dixie J. Statham, Kathleen K. Bucholz, Pamela A. F. Madden, Andrew C. Heath, Nicholas G. Martin, and Elliot C. Nelson. Common Heritable Contributions to Low-Risk Trauma, High-Risk Trauma, Posttraumatic Stress Disorder, and Major Depression. *Archives of general psychiatry*, 69(3):293–299, March 2012.

[254] J M Schildkraut, N Risch, and W D Thompson. Evaluating genetic association among ovarian, breast, and endometrial cancer: Evidence for a breast/ovarian cancer relationship. *American Journal of Human Genetics*, 45(4):521–529, October 1989.

[255] Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, July 2014.

[256] Aswin Sekar, Allison R. Bialas, Heather de Rivera, Avery Davis, Timothy R. Hammond, Nolan Kamitaki, Katherine Tooley, Jessy Presumey, Matthew Baum, Vanessa Van Doren, Giulio Genovese, Samuel A. Rose, Robert E. Handsaker, Schizophrenia

Working Group of the Psychiatric Genomics Consortium, Mark J. Daly, Michael C. Carroll, Beth Stevens, and Steven A. McCarroll. Schizophrenia risk from complex variation of complement component 4. *Nature*, 530(7589):177–183, February 2016.

[257] Carlo Selmi, Marlyn J. Mayo, Nancy Bach, Hiromi Ishibashi, Pietro Invernizzi, Robert G. Gish, Stuart C. Gordon, Harlan I. Wright, Bruce Zweiban, Mauro Podda, and M. Eric Gershwin. Primary biliary cirrhosis in monozygotic and dizygotic twins: Genetics, epigenetics, and environment. *Gastroenterology*, 127(2):485–492, August 2004.

[258] Kenji Shibuya and Eiji Yano. Regression analysis of trends in mortality from hepatocellular carcinoma in Japan, 1972-2001. *International Journal of Epidemiology*, 34(2):397–402, April 2005.

[259] Claudia Tamar Silva, Jan A. Kors, Najaf Amin, Abbas Dehghan, Jacqueline C. M. Witteman, Rob Willemsen, Ben A. Oostra, Cornelia M. van Duijn, and Aaron Isaacs. Heritabilities, proportions of heritabilities explained by GWAS findings, and implications of cross-phenotype effects on PR interval. *Human Genetics*, 134(11-12):1211–1219, November 2015.

[260] Jodie L. Simpson, Rodney Scott, Michael J. Boyle, and Peter G. Gibson. Inflammatory subtypes in asthma: Assessment and identification using induced sputum. *Respirology*, 11(1):54–61, January 2006.

[261] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810, February 2016.

[262] Valérie Siroux, Lydiane Agier, and Rémy Slama. The exposome concept: A challenge and a potential driver for environmental health research. *European Respiratory Review: An Official Journal of the European Respiratory Society*, 25(140):124–129, June 2016.

[263] Masoud Soheilian, Mohammad-Mehdi Sadoughi, Mehdi Ghajarnia, Mohammad H. Dehghan, Shahin Yazdani, Hassan Behboudi, Arash Anisian, and Gholam A. Peyman. Prospective randomized trial of trimethoprim/sulfamethoxazole versus pyrimethamine and sulfadiazine in the treatment of ocular toxoplasmosis. *Ophthalmology*, 112(11):1876–1882, November 2005.

[264] Nadia Solovieff, Chris Cotsapas, Phil H. Lee, Shaun M. Purcell, and Jordan W. Smoller. Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, July 2013.

[265] D. Sorensen and D. Gianola. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. New York: Springer, 2002.

[266] Doug Speed, Terence J. O'Brien, Aarno Palotie, Kirill Shkura, Anthony G. Marson, David J. Balding, and Michael R. Johnson. Describing the genetic architecture of epilepsy through heritability analysis. *Brain: A Journal of Neurology*, 137(Pt 10):2680–2689, October 2014.

[267] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, October 2002.

[268] Henriët. Springelkamp, René Höhn, Aniket Mishra, Pirro G. Hysi, Chiea-Chuen Khor, Stephanie J. Loomis, Jessica N. Cooke Bailey, Jane Gibson, Gudmar Thorleifsson, Sarah F. Janssen, Xiaoyan Luo, Wishal D. Ramdas, Eranga Vithana, Monisha E. Nongpiur, Grant W. Montgomery, Liang Xu, Jenny E. Mountain, Puya Gharahkhani, Yi Lu, Najaf Amin, Lennart C. Karssen, Kar-Seng Sim, Elisabeth M. van Leeuwen, Adriana I. Iglesias, Virginie J. M. Verhoeven, Michael A. Hauser, Seng-Chee Loon, Dominiek D. G. Despriet, Abhishek Nag, Cristina Venturini, Paul G. Sanfilippo, Arne Schillert, Jae H. Kang, John Landers, Fridbert Jonasson, Angela J. Cree, Leonieke M. E. van Koolwijk, Fernando Rivadeneira, Emmanuelle Souzeau, Vesteinn Jonsson, Geeta Menon, Robert N. Weinreb, Paulus T. V. M. de Jong, Ben A. Oostra, André G. Uitterlinden, Albert Hofman, Sarah Ennis, Unnur Thorsteinsdottir, Kathryn P. Burdon, Timothy D. Spector, Alireza Mirshahi, Seang-Mei Saw, Johannes R. Vingerling, Yik-Ying Teo, Jonathan L. Haines, Roger C. W. Wolfs, Hans G. Lemij, E-Shyong Tai, Nomdo M. Jansonius, Jost B. Jonas, Ching-Yu Cheng, Tin Aung, Ananth C. Viswanathan, Caroline C. W. Klaver, Jamie E. Craig, Stuart Macgregor, David A. Mackey, Andrew J. Lotery, Kari Stefansson, Arthur A. B. Bergen, Terri L. Young, Janey L. Wiggs, Norbert Pfeiffer, Tien-Yin Wong, Louis R. Pasquale, Alex W. Hewitt, Cornelia M. van Duijn, and Christopher J. Hammond. Meta-analysis of genome-wide association studies identifies novel loci that influence cupping and the glaucomatous process. *Nature Communications*, 5, September 2014.

[269] Eli A. Stahl, Soumya Raychaudhuri, Elaine F. Remmers, Gang Xie, Stephen Eyre, Brian P. Thomson, Yonghong Li, Fina A. S. Kurreeman, Alexandra Zhernakova, Anne Hinks, Candace Guiducci, Robert Chen, Lars Alfredsson, Christopher I. Amos, Kristin G. Ardlie, BIRAC Consortium, Anne Barton, John Bowes, Elisabeth Brouwer, Noel P. Burtt, Joseph J. Catanese, Jonathan Coblyn, Marieke J. H. Coenen, Karen H. Costenbader, Lindsey A. Criswell, J. Bart A. Crusius, Jing Cui, Paul I. W. de Bakker, Philip L. De Jager, Bo Ding, Paul Emery, Edward Flynn, Pille Harrison, Lynne J. Hocking, Tom W. J. Huizinga, Daniel L. Kastner, Xiayi Ke, Annette T. Lee, Xiangdong Liu, Paul Martin, Ann W. Morgan, Leonid Padyukov, Marcel D. Posthumus, Timothy R. D. J. Radstake, David M. Reid, Mark Seielstad, Michael F. Seldin, Nancy A. Shadick, Sophia Steer, Paul P. Tak, Wendy Thomson, Annette H. M. van der Helm-van Mil, Irene E. van der Horst-Bruinsma, C. Ellen van der Schoot, Piet L. C. M. van Riel, Michael E. Weinblatt, Anthony G. Wilson, Gert Jan Wolbink, B. Paul Wordsworth, YEAR Consortium, Cisca Wijmenga, Elizabeth W. Karlson, Rene E. M. Toes, Niek de Vries, Ann B. Begovich, Jane Worthington, Katherine A. Siminovitch,

Peter K. Gregersen, Lars Klareskog, and Robert M. Plenge. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics*, 42(6):508–514, June 2010.

[270] StataCorp. Stata Statistical Software: Release 13. 2013.

[271] Eduardo Steidl, Carla Simone Ribeiro, Bruna Franciele Gonçalves, Natália Fernandes, Vívian Antunes, and Renata Mancopes. Relationship between Dysphagia and Exacerbations in Chronic Obstructive Pulmonary Disease: A Literature Review. *International Archives of Otorhinolaryngology*, 19(1):74–79, January 2015.

[272] Murray B. Stein, Chia-Yen Chen, Robert J. Ursano, Tianxi Cai, Joel Gelernter, Steven G. Heeringa, Sonia Jain, Kevin P. Jensen, Adam X. Maihofer, Colter Mitchell, Caroline M. Nievergelt, Matthew K. Nock, Benjamin M. Neale, Renato Polimanti, Stephan Ripke, Xiaoying Sun, Michael L. Thomas, Qian Wang, Erin B. Ware, Susan Borja, Ronald C. Kessler, and Jordan W. Smoller. Genome-wide Association Studies of Posttraumatic Stress Disorder in 2 Cohorts of US Army Soldiers. *JAMA Psychiatry*, 73(7):695–704, July 2016.

[273] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toksöz, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E. Wanker. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122(6):957–968, September 2005.

[274] Eileen Stillwaggon, Christopher S. Carrier, Mari Sautter, and Rima McLeod. Maternal serologic screening to prevent congenital toxoplasmosis: A decision-analytic economic model. *PLoS neglected tropical diseases*, 5(9):e1333, September 2011.

[275] P. Sturmey and V. James. Administrative prevalence of autism in the Texas school system. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(6):621, June 2001.

[276] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*, 12(3):e1001779, March 2015.

[277] Mervyn Susser and Zena Stein. Civilization and peptic ulcer. *International Journal of Epidemiology*, 31(1):13–17, January 2002.

[278] Silpa Suthram, Joel T. Dudley, Annie P. Chiang, Rong Chen, Trevor J. Hastie, and Atul J. Butte. Network-Based Elucidation of Human Disease Similarities Reveals

Common Functional Modules Enriched for Pluripotent Drug Targets. *PLoS Comput Biol*, 6(2):e1000662, February 2010.

[279] Yi Hang Tay, Milawaty Nurjono, and Jimmy Lee. Increased Framingham 10-year CVD risk in Chinese patients with schizophrenia. *Schizophrenia Research*, 147(1):187–192, June 2013.

[280] R Core Team. R: A Language and Environment for Statistical Computing. 2013.

[281] the EArly Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics*, 47(12):1449–1456, December 2015.

[282] Paul Thompson, Syed A. Iqbal, John McNaught, and Sophia Ananiadou. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10:349, 2009.

[283] Carl Peter Thunberg, Peter Ulrik Berg, Carl Fredrik Blumenberg, Nils Gustaf Bodin, Pehr Branström, Claës Fredrik Hornstedt, Carl Fredrik Lexow, Johan Gustaf Lodin, Claus Erik Mellerborg, Carl Henrik Salberg, Andreas Gustaf Salmenius, Carl Fredrik Sjöbeck, Gustaf Erik Sörling, Gabriel Tobias Ström, Erik Carl Trafvenfeldt, Conrad Wallenius, and Samuel Wallner. *Nova Genera Plantarum...* apud. J. Edman [etc.,, Upsaliæ :, 1781.

[284] Beata Tick, Patrick Bolton, Francesca Happé, Michael Rutter, and Frühling Rijsdijk. Heritability of autism spectrum disorders: A meta-analysis of twin studies. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 57(5):585–595, May 2016.

[285] Jorim J. Tielbeek, Sarah E. Medland, Beben Benyamin, Enda M. Byrne, Andrew C. Heath, Pamela A. F. Madden, Nicholas G. Martin, Naomi R. Wray, and Karin J. H. Verweij. Unraveling the Genetic Etiology of Adult Antisocial Behavior: A Genome-Wide Association Study. *PLoS ONE*, 7(10), October 2012.

[286] D. Torre, S. Casari, F. Speranza, A. Donisi, G. Gregis, A. Poggio, S. Ranieri, A. Orani, G. Angarano, F. Chiodo, G. Fiori, and G. Carosi. Randomized trial of trimethoprim-sulfamethoxazole versus pyrimethamine-sulfadiazine for therapy of toxoplasmic encephalitis in patients with AIDS. Italian Collaborative Study Group. *Antimicrobial Agents and Chemotherapy*, 42(6):1346–1349, June 1998.

[287] E. Fuller Torrey, John J. Bartko, Zhao-Rong Lun, and Robert H. Yolken. Antibodies to Toxoplasma gondii in patients with schizophrenia: A meta-analysis. *Schizophrenia Bulletin*, 33(3):729–736, May 2007.

[288] M. F. Trulson, R. E. Clancy, W. J. Jessop, R. W. Childers, and F. J. Stare. COMPARISONS OF SIBLINGS IN BOSTON AND IRELAND. *Journal of the American Dietetic Association*, 45:225–229, September 1964.

[289] C. Tysk, E. Lindberg, G. Järnerot, and B. Flodérus-Myrhed. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut*, 29(7):990–996, July 1988.

[290] Nian-Sheng Tzeng, Yung-Ho Hsu, Shinn-Ying Ho, Yu-Ching Kuo, Hua-Chin Lee, Yun-Ju Yin, Hong-An Chen, Wen-Liang Chen, William Cheng-Chung Chu, and Hui-Ling Huang. Is schizophrenia associated with an increased risk of chronic kidney disease? A nationwide matched-cohort study. *BMJ Open*, 5(1):e006777, January 2015.

[291] United States Census Bureau. *Children by Presence and Type of Parent(s), Race, and Hispanic Origin:2007-2011*. 2007-2011. Online; accessed 29 Feb 2016.

[292] US United States Environmental Protection Agency. Birth Defects Prevalence and Mortality EPA. `http://cfpub.epa.gov/eroe/index.cfm?fuseaction=detail.viewInd\&lv=list.listbyalpha\&r=239796\&subtop=381`.

[293] Data Integration Division US Census Bureau. Population Estimates. `http://www.census.gov/popest/data/historical/2010s/vintage_2011/`.

[294] Jose M. Valderas, Barbara Starfield, Bonnie Sibbald, Chris Salisbury, and Martin Roland. Defining Comorbidity: Implications for Understanding Health and Health Services. *Annals of Family Medicine*, 7(4):357–363, July 2009.

[295] Tanya van de Water, Sharain Suliman, and Soraya Seedat. Gender and cultural issues in psychiatric nosological classification systems. *CNS spectrums*, 21(4):334–340, August 2016.

[296] Karla C. Van Meter, Lasse E. Christiansen, Lora D. Delwiche, Rahman Azari, Tim E. Carpenter, and Irva Hertz-Picciotto. Geographic distribution of autism in California: A retrospective birth cohort analysis. *Autism Research: Official Journal of the International Society for Autism Research*, 3(1):19–29, February 2010.

[297] MJ van Rijn, AF Schut, YS Aulchenko, J Deinum, and et al. Heritability of blood pressure traits and the genetic contribution to blood pressure variance explained by four blood-pressure-related genes. *J Hypertens*, (25):565–57, 2007.

[298] Shashaank Vattikuti, Juen Guo, and Carson C. Chow. Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits. *PLOS Genetics*, 8(3):e1002637, March 2012.

[299] Philip A. Verhoef, Michael G. Constantinides, Benjamin D. McDonald, Joseph F. Urban, Anne I. Sperling, and Albert Bendelac. Intrinsic functional defects of type 2 innate lymphoid cells impair innate allergic inflammation in promyelocytic leukemia zinc finger (PLZF)-deficient mice. *The Journal of Allergy and Clinical Immunology*, 137(2):591–600.e1, February 2016.

[300] Carl von Linné and Eng. Lichfield. *The Families of Plants, with Their Natural Characters, According to the Number, Figure, Situation, and Proportion of All the Parts of Fructification.* Genera plantarum.English. Printed by J. Jackson, Lichfield, 1787.

[301] Carl Magnus Wahlgren and Patrik K. E. Magnusson. Genetic influences on peripheral arterial disease in a twin population. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 31(3):678–682, March 2011.

[302] M. Wallon, F. Peyron, C. Cornu, S. Vinault, M. Abrahamowicz, C. Bonithon Kopp, and C. Binquet. Congenital toxoplasma infection: Monthly prenatal screening decreases transmission rate and improves clinical outcome at age 3 years. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 56(9):1223–1231, May 2013.

[303] Kanix Wang, Hallie Gaitsch, Hoifung Poon, Nancy J. Cox, and Andrey Rzhetsky. Classification of common human diseases derived from shared genetic and environmental determinants. *Nature Genetics*, advance online publication, August 2017.

[304] Xiaoyan Wang, Amy Chused, Noémie Elhadad, Carol Friedman, and Marianthi Markatou. Automated Knowledge Acquisition from Clinical Narrative Reports. *AMIA Annual Symposium Proceedings*, 2008:783–787, 2008.

[305] Casey T. Weaver, Charles O. Elson, Lynette A. Fouser, and Jay K. Kolls. The Th17 pathway and inflammatory diseases of the intestines, lungs, and skin. *Annual Review of Pathology*, 8:477–512, January 2013.

[306] Wei-Qi Wei and Joshua C. Denny. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Medicine*, 7:41, 2015.

[307] Wei-Qi Wei, Pedro L. Teixeira, Huan Mo, Robert M. Cronin, Jeremy L. Warner, and Joshua C. Denny. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association*, 23(e1):e20–e27, April 2016.

[308] Sally E. Wenzel, Lawrence B. Schwartz, Esther L. Langmack, Janet L. Halliday, John B. Trudeau, Robyn L. Gibbs, and Hong Wei Chu. Evidence That Severe Asthma Can Be Divided Pathologically into Two Inflammatory Subtypes with Distinct Physiologic and Clinical Characteristics. *American Journal of Respiratory and Critical Care Medicine*, 160(3):1001–1008, September 1999.

[309] WHO. ICD International Classification of Diseases. `http://www.who.int/classifications/icd/en/`, 2017.

[310] Gonneke Willemsen, Kirsten J. Ward, Christopher G. Bell, Kaare Christensen, Jocelyn Bowden, Christine Dalgård, Jennifer R. Harris, Jaakko Kaprio, Robert Lyle, Patrik K. E. Magnusson, Karen A. Mather, Juan R. Ordoňana, Francisco Perez-Riquelme, Nancy L. Pedersen, Kirsi H. Pietiläinen, Perminder S. Sachdev, Dorret I. Boomsma, and Tim Spector. The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies*, 18(6):762–771, December 2015.

[311] Jesse W. Williams, Melissa Y. Tjota, Bryan S. Clay, Bryan Vander Lugt, Hozefa S. Bandukwala, Cara L. Hrusch, Donna C. Decker, Kelly M. Blaine, Bethany R. Fixsen, Harinder Singh, Roger Sciammas, and Anne I. Sperling. Transcription factor IRF4 drives dendritic cells to promote Th2 differentiation. *Nature Communications*, 4:2990, 2013.

[312] Andrea N. Witwer and Luc Lecavalier. Examining the Validity of Autism Spectrum Disorder Subtypes. *Journal of Autism and Developmental Disorders*, 38(9):1611–1624, October 2008.

[313] Nicole I. Wolf, Gajja S. Salomons, Richard J. Rodenburg, Petra J. W. Pouwels, Jolanda H. Schieving, Terry G. J. Derks, Johanna M. Fock, Patrick Rump, Daphne M. van Beek, Marjo S. van der Knaap, and Quinten Waisfisz. Mutations in RARS cause hypomyelination. *Annals of Neurology*, 76(1):134–139, July 2014.

[314] Brian Wong. Central-Nervous-System Toxoplasmosis in Homosexual Men and Parenteral Drug Abusers. *Annals of Internal Medicine*, 100(1):36, January 1984.

[315] Prescott G. Woodruff, Barmak Modrek, David F. Choy, Guiquan Jia, Alexander R. Abbas, Almut Ellwanger, Joseph R. Arron, Laura L. Koth, and John V. Fahy. T-helper Type 2–driven Inflammation Defines Major Subphenotypes of Asthma. *American Journal of Respiratory and Critical Care Medicine*, 180(5):388–395, September 2009.

[316] Naomi R. Wray, Jian Yang, Michael E. Goddard, and Peter M. Visscher. The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling. *PLOS Genetics*, 6(2):e1000864, February 2010.

[317] S. Wright. Systems of Mating. I. the Biometric Relations between Parent and Offspring. *Genetics*, 6(2):111–123, March 1921.

[318] Charley Xia, Carmen Amador, Jennifer Huffman, Holly Trochet, Archie Campbell, David Porteous, Generation Scotland, Nicholas D. Hastie, Caroline Hayward, Veronique Vitart, Pau Navarro, and Chris S. Haley. Pedigree- and SNP-Associated Genetics and Recent Environment are the Major Contributors to Anthropometric and Cardiometabolic Trait Variation. *PLOS Genetics*, 12(2):e1005804, February 2016.

[319] Junping Yan, Bo Huang, Guochen Liu, Bin Wu, Shiguang Huang, Huanqin Zheng, Jilong Shen, Zhao-Rong Lun, Yong Wang, and Fangli Lu. Meta-analysis of prevention and treatment of toxoplasmic encephalitis in HIV-infected patients. *Acta Tropica*, 127(3):236–244, September 2013.

[320] L. Yao, J. W. Berman, S. M. Factor, and F. D. Lowy. Correlation of histopathologic and bacteriologic changes with cytokine expression in an experimental murine model of bacteremic Staphylococcus aureus infection. *Infection and Immunity*, 65(9):3889–3895, January 1997.

[321] Yee Guan Yap and A. John Camm. Drug induced QT prolongation and torsades de pointes. *Heart (British Cardiac Society)*, 89(11):1363–1372, November 2003.

[322] Dietmar M. W. Zaiss, William C. Gause, Lisa C. Osborne, and David Artis. Emerging functions of amphiregulin in orchestrating immunity, inflammation, and tissue repair. *Immunity*, 42(2):216–226, February 2015.

[323] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L. Price. Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLOS Genetics*, 9(5):e1003520, May 2013.

[324] Qing T. Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N. Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6:30, July 2006.

[325] Jie Zheng, A. Mesut Erzurumluoglu, Benjamin L. Elsworth, John P. Kemp, Laurence Howe, Philip C. Haycock, Gibran Hemani, Katherine Tansey, Charles Laurin, Early Genetics Consortium, Lifecourse Epidemiology (EAGLE) Eczema, Beate St Pourcain, Nicole M. Warrington, Hilary K. Finucane, Alkes L. Price, Brendan K. Bulik-Sullivan, Verneri Anttila, Lavinia Paternoster, Tom R. Gaunt, David M. Evans, and Benjamin M. Neale. LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, page btw613, September 2016.

[326] Jin J. Zhou, Michael H. Cho, Peter J. Castaldi, Craig P. Hersh, Edwin K. Silverman, and Nan M. Laird. Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *American Journal of Respiratory and Critical Care Medicine*, 188(8):941–947, October 2013.

[327] George Kingsley Zipf. *The Psycho-Biology of Language; an Introduction to Dynamic Philology*. Houghton Mifflin company, Boston, 1935.

[328] Or Zuk, Eliana Hechter, Shamil R. Sunyaev, and Eric S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, January 2012.