UNIVERSITY OF CALIFORNIA

Los Angeles

Computational methods for disease diagnosis and understanding the genetics of complex traits

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Lisa Gai

2021

ABSTRACT OF THE DISSERTATION

Los Angeles

Computational methods for disease diagnosis and understanding

the genetics of complex traits

by

Lisa Gai

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2021

Professor Eleazar Eskin, Chair

An ever increasing wealth of biological data has become available in recent years, and with it, the potential to understand complex traits and extract disease relevant information from these many forms of data through computational methods. Understanding the genetic architecture behind complex traits can help us understand disease risk and adverse drug reactions, and to guide the development of treatment strategies. Many variants identified by genome-wide association studies (GWAS) have been found to affect multiple traits, either directly or through shared pathways. Analyzing multiple traits at once can increase power to detect shared variant effects from publicly available GWAS summary statistics. Use of multiple traits may also improve accuracy when estimating variant effects, which can be used in polygenic scores to stratify individuals by disease risk. This dissertation presents a

method, CONFIT, for combining GWAS in multiple traits for variant discovery, and explores a few potential multi-trait methods for estimating polygenic scores. Computational methods can also be used to identify patients already suffering from disease who would benefit from treatment. Towards this end, this dissertation also presents work on deep learning to detect patients with orbital disease from image data with high accuracy and recall.

The dissertation of Lisa Gai is approved.

Jason Ernst

Sriram Sankararaman

Wei Wang

Eleazar Eskin, Committee Chair

University of California, Los Angeles

2021

Contents

List of Figures

List of Tables

Vita

2010 - 2014 B.S. in Mathematical and Computation Biology with Honors in Computer Science, Harvey Mudd College, Claremont, CA

Publications

- Dat Duong, **Lisa Gai**, Ankith Uppunda, Don Le, Eleazar Eskin, Jessica Jingyi Li, Kai-Wei Chang. Annotating Gene Ontology terms for protein sequences with the Transformer model. *bioRxiv preprint* (2020).

- Dat Duong, Ankith Uppunda, **Lisa Gai**, Chelsea Jui-Ting Ju, James Zhang, Muhao Chen, Eleazar Eskin, Jessica Jingyi Li, Kai-Wei Chang. Evaluating Representations for Gene Ontology Terms. *bioRxiv preprint* (2020).

- **Lisa Gai**, Eleazar Eskin. Finding associated variants in genome-wide association studies on multiple traits. *Bioinformatics* (2018).

- Dat Duong, **Lisa Gai**, Sagi Snir, Eun Yong Kang, Buhm Han, Jae Hoon Sul, Eleazar Eskin. Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. *Bioinformatics* (2017).

In preparation

– **Lisa Gai**[†], Justin Karlin[†], Nathan LaPierre, Kayla Danesh, Justin Farajzadeh, Bea Palileo, Kodi Taraszka, Jie Zheng, Wei Wang, Eleazar Eskin, Daniel Rootman. Deep neural network guided detection of thyroid eye disease from external photos.

– **Lisa Gai**, Jingyuan Fu, Camille Huang, Sriram Sankararaman, Eleazar Eskin. Evaluating models for variant effects across multiple traits using using summary statistics.

Fellowships

– Microsoft Research Graduate Women's Scholarship (2015-2016)

– UCLA Eugene V. Cota-Robles Fellowship (2014-2018)

Chapter 1

Introduction

1.1 Motivation

Computational methods can help us make sense of the vast span of information in the human genome. Understanding the genetic architecture behind complex traits can help us predict traits such as disease risk and adverse drug reactions, and to guide the development of treatment strategies. In a more direct way, computational methods can also be used to identify patients already suffering from disease who would benefit from such treatment.

Both aims motivated this thesis work, which can be roughly split into statistical methods for combining genome-wide association data from multiple traits, and deep learning for disease diagnosis.

1.2 Combining genetic association studies on multiple traits

In the first part of my PhD research, I focused on two projects, (1) CONFIT, a method to increasing power in association tests using study results from different traits, and (2) a comparison of multi-trait models for effect sizes that allow for varying degrees of correlation across the genome. Prior to that, I assisted in the development of RECOV, a random effects model that allows for correlation between studies and was applied to model loci affecting gene expression across related tissues (Duong *et al.*, 2017). This experience inspired my foray into multi-trait projects.

1.2.1 Background on genetic association studies

Before talking about these projects in more detail, I provide background on the type of data used. Genome-wide association studies (GWAS) are used to find genetic variation correlated with a phenotype such as height, gene expression level, or disease status. Typically they focus on single-nucelotide polymorphisms (SNPs) where one nucleotide is swapped for another at a single position in the genome in some portion of the population. Each individual may have one of the swapped or unswapped versions (alleles) at each choromosome. The alleles may be labeled as the major and minor alleles based on which is more common in the population, or the derived and ancestral allele based on which originated first.

To conduct a GWAS, genotype and phenotype data is collected for many individuals – perhaps a hundred or so for a brain tissue phenotype which is hard to collect, or hundreds of thousands for concerted biobank efforts. Then for each position in the genome with a SNP,

in essence a linear or logistic regression is performed with the allele count (how many copies of the minor allele each person has) as the independent variable and the phenotype as the dependent varible. This produces an estimate of the effect size and standard error for each SNP, which are known as the summary statistics. These summary statistics often readily available in contrast to the original genotype and phenotype data, due to privacy concerns and requiring much greater computing resources to store and use compared to summary statistics.

GWAS are not without their flaws, as addressed by numerous papers in the literature that improve upon GWAS. However, their simplicity and ease of use, as well as the convenience of sharing summary statistics, have made them an enduring feature of genetics research.

1.2.2 Association testing using GWAS from multiple traits

One limitation of GWAS is that SNP effect sizes may be very small and obtaining sample sizes large enough to obtain sufficient statistical power infeasible. Combining results from multiple studies can increase power. Many variants identified by genome-wide association studies (GWAS) have been found to affect multiple traits, a phenomenon known as pleiotropy. Analyzing multiple traits at once can increase power to detect and estimate shared variant effects. For combining multiple GWAS on the same phenotype, one can use meta-analysis increase power, as in (Postmus *et al.*, 2016; Nikpay *et al.*, 2015; Berndt *et al.*, 2016). However, meta-analysis methods may have decreased power when effects are unique to a particular study or subset of studies in the analysis, making them unsuitable for combining studies in different traits. In this project, we developed CONFIT, a method for association testing from

summary statistics in multiple traits. CONFIT allows different "configurations" of non-zero effects amongst traits, i.e. effects may be present in different subsets of traits used in the analysis. The CONFIT test statistic is a likelihood ratio averaged over these configurations, where the contribution of each configuration is weighted by a prior estimated from the data.

Using simulated data, we show that CONFIT properly controls false positive rate and increases power when traits are correlated. When applied to a set of five metabolic traits measured in the North Finland Birth Cohort, CONFIT was able to discover eight of nine GWAS-significant loci, as well as two additional loci that were replicated in larger studies. We also applied CONFIT to four metabolic traits in the UK Biobank, and discovered 44 novel loci related to both high cholesterol and use of cholesterol medication. This work was published in *Bioinformatics* in June 2018 (Gai and Eskin, 2018) and is also presented in Chapter 2.

1.2.3 Estimating effects and computing polygenic scores using GWAS from multiple traits

Polygenic scores (PGS) have emerged as a promising tool for disease risk assessment (Shieh *et al.*, 2016; Logue *et al.*, 2018), which can inform the need for early screening or lifestyle changes. PGS estimate an individual's phenotype based on their genetics by aggregating the effects from many variants, typically using effect sizes estimated from GWAS on the trait of interest. Improving the accuracy of the estimates would in turn improve the predictive power of the PGS. Many pairs of traits exhibit genetic correlation, i.e. their effect sizes are correlated, and exhibit significant genetic correlation even in the absence of any significantly

associated loci (Bulik-Sullivan *et al.*, 2015a; Shi *et al.*, 2017). Modeling effects across multiple traits jointly can improve estimate shared variant effects by using this genetic correlation.

In this project, we consider several multi-trait models for effect sizes and assess their performance on a variety of traits from the UK Biobank. We find that multi-trait PGS methods can increase PGS accuracy, particularly when the GWAS for the trait of interest is small and it has strong genetic with other traits in the analysis, but can perform worse than the original GWAS in other scenarios. We also find that models allowing for variable patterns of genetic correlation across the genome do not consistently offer performance benefits over a method that assumes uniform genetic correlation (Turley *et al.*, 2018), or against GWAS itself. This work is presented in Chapter 3.

1.3 Identifying an orbital disease from image data using neural networks

The second part of my PhD work was on deep learning methods for biology and medicine. My primary work here was to develop and evaulate an ensemble of neural networks model for identifying patients with thyroid eye disease from digital photographs, presented in Chapter 4and summarized below. I also worked on two projects led by Dat Duong on applying deep learning to Gene Ontology (GO) terms: a comparison of graph convolutional networks, ELMo, and BERT for producing embedding of GO terms (Duong *et al.*, 2020b), and GO annotation based on the transformer framework (GOAT), a method for predicting GO terms for a protein sequence (Duong *et al.*, 2020a).

Thyroid eye disease (TED) is an autoimmune disease causing inflammation of the orbital tissues (Bahn, 2010). Severe cases may result in disfigurement, chronic eye pain, and misalignment of the eyes (Bahn, 2010; Sabini *et al.*, 2017). Timely treatment can reduce the degree of inflammation and severity of final outcome. However, the time from a patient's initial hospital visit to diagnosis often takes months or even years (Mellington *et al.*, 2017; Estcourt *et al.*, 2009).

Deep learning has shown excellent performance in many areas of image recognition, and holds great promise in the medical domain. In ophthalmology, it has been primarily been applied to identify conditions from fundus (retina images), optical coherence tomography (OCT), or computed tomography (CT).

In this project, we developed a deep learning-based classifier to identify TED from simple digital photos of the face, with the goal of helping primary care physicians or even patients themselves to quickly and effectively screen for TED in the future. The classifier is an ensemble of neural networks, trained on 1,252 control images and 692 TED images obtained from UCLA Stein Eye Institute (SEI) clinical data. It achieved an overall accuracy of 89% and recall rate of 93% on 46 held out patient images from SEI. When applied to 122 patient images from a separate clinical practice wholly unseen during training, it achieved a recall rate of 92% with higher recall for more severe cases. The full details of this work are given in Chapter 4.

Chapter 2

Finding associated variants in genome-wide association studies on multiple traits

This work was published in Bioinformatics in 2018.

2.1 Introduction

Over the past few decades, genome wide association studies (GWAS) have found numerous genetic variants associated with phenotypic variation (McCarthy *et al.*, 2008; Dorn and Cresci, 2009; Eskin, 2015). These phenotypes include a wide range of diseases and medically relevant traits such as heart disease (Dorn and Cresci, 2009; Lee *et al.*, 2013; Nikpay *et al.*, 2015), cholesterol level (Postmus *et al.*, 2016), and depression (Cai *et al.*, 2015; Hyde *et al.*, 2016), among others. In some cases, variants have been found to affect multiple

traits, a phenomenon known as pleiotropy (Andreassen *et al.*, 2014). For example, multiple psychiatric disorders, immune diseases, and nervous system phenotypes have been found to share causal variants (Solovieff *et al.*, 2013; Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; Chen *et al.*, 2016; Chesler *et al.*, 2005; Zeggini and Ioannidis, 2009). Variants associated with disease have also been found to be associated with tissue-specific gene expression phenotypes (Liu *et al.*, 2016). Considering multiple traits at once may increase power to detect variant effects when there is pleiotropy.

One approach to combine information from different studies is to apply meta-analysis. Meta-analysis methods are often used in GWAS to combine results from different studies on the same trait to increase power (Postmus *et al.*, 2016; Nikpay *et al.*, 2015; Berndt *et al.*, 2016). Intuitively, one can effectively increase the sample size by pooling summary statistics from multiple small studies, which also have the benefit of being more readily obtainable compared to individual level data. The two classic versions of meta-analysis are fixed effects (FE) meta-analysis and random effects (RE) meta-analysis (Fleiss, 1993). In the FE model, a variant is assumed to have the same effect in each study, which is only realistic if all studies in the meta-analysis measure the same phenotype in the same population. If instead the true effect size differs between studies, we say there is heterogeneity. The random effects (RE) model allows for heterogeneity by assuming study-specific effect sizes are drawn independently from a normal distribution. The binary effects (BE) model also allows for heterogeneity (Han and Eskin, 2012). In BE meta-analysis, a variant may either have an effect of fixed size or no effect in each study (Han and Eskin, 2012). A variant's configuration of effects across traits may then be expressed as binary vector with entries indicating whether or not the effect is zero for each trait.

However, it is problematic to directly apply meta-analysis to combine studies that analyze different traits for a number of reasons. First, some traits share many causal variants while others share very few. Existing meta-analysis methods do not allow for varying degrees of shared variants between traits, and combining unrelated traits in a meta-analysis may actually decrease power compared to independent analysis of such traits. Second, a variant that affects one trait may have no effect in a different trait. While RE meta-analysis and related methods allow for differences in effect size between studies, such methods inherently assume an effect is present in all studies in the meta-analysis. Finally, studies may share individuals across traits. For example, data on several traits may be collected from the same cohort of individuals. Meta-analysis techniques assume that the studies are independent, but this only holds if the studies are performed on non-overlapping individuals.

In this paper, we present CONFIT, a novel meta-analysis method for multiple traits that addresses these shortcomings. CONFIT estimates the degree of shared effects between traits from the data using GWAS summary statistics, then uses these estimates to analyze multiple traits while allowing effects to be present in only a subset of the traits. CONFIT is inspired by the existence of pleiotropy and its potential to increase power to detect variants that affect multiple traits. Unlike traditional meta-analysis methods, CONFIT is designed to combine GWAS on different traits and does not assume a particular relationship between the different traits. Our test statistic is a likelihood ratio averaged over many models, where each model assumes the variant to have non-zero effect in a particular subset of traits and is weighted by a prior estimated from the data.

We tested CONFIT and show it has increased power compared to multiple independent GWAS in simulated data when variants have effect in multiple traits. We also show CONFIT

accounts for correlated effect size estimates from overlapping individuals between studies. We then demonstrate that CONFIT finds unique loci when combining studies on multiple traits using the North Finland Birth Cohort (NFBC) data set and the UK Biobank data set. CONFIT has many potential applications due to the vast variety of GWAS data sets available.

2.2 Methods

2.2.1 Finding associated variants in one trait using a genome-wide association study (GWAS)

We now describe how to test a variant v for association in a trait t using a GWAS. Let $\boldsymbol{g}_{vt}$ be the vector of genotype values in n_t individuals collected in the study for trait t. Denote entry j in $\boldsymbol{g}_{vt}$ as $g_{vt,j}$, which corresponds to the genotype of the jth individual in study t, i.e. the number of copies of variant v they possess. Thus $g_{vt,j} \in \{0, 1, 2\}$. Let $\boldsymbol{x}_{vt}$ be the vector of standardized genotype values in study t. In other words, $\boldsymbol{x}_{vt}$ is obtained by mean-centering and scaling $\boldsymbol{g}_{vt}$ to have a sample variance of 1.

Let $\boldsymbol{y}_t$ be the vector of phenotype values in n_t individuals for trait t. Assume $\boldsymbol{y}_t$ has been centered to have mean 0. Given $\boldsymbol{x}_{vt}, \boldsymbol{y}_t$, GWAS assumes the linear model

$$\boldsymbol{y}_t = \beta_{vt}\boldsymbol{x}_{vt} + \boldsymbol{e}_t \tag{2.1}$$

where β_{vt} is the effect of v on trait t and $\boldsymbol{e}_t \sim N(0, \sigma_e^2\boldsymbol{I})$ is gaussian noise (Eskin, 2015). The magnitude of β_{vt} indicates how predictive v is. One then finds the estimated effect $\hat{\beta}_{vt}$

by linear regression. The solution given by ordinary least squares (OLS) is

$$\hat{\beta}_{vt} = (\boldsymbol{x}_{vt}^{\intercal}\boldsymbol{x}_{vt})^{-1}\boldsymbol{x}_{vt}^{\intercal}\boldsymbol{y}_k \tag{2.2}$$

where

$$\hat{\beta}_{vt} \sim N(\beta_{vt}, (\boldsymbol{x}_{vt}^{\intercal}\boldsymbol{x}_{vt})^{-1}\sigma_e^2) \tag{2.3}$$

Since σ_e^2 is unknown, we estimate it as $\hat{\sigma}_e^2 = \frac{1}{n_t - 1}||\boldsymbol{y}_t - \hat{\beta}_{vt}\boldsymbol{x}_{vt}||_2^2$. Let $d_{vt} = (\boldsymbol{x}_{vt}^{\intercal}\boldsymbol{x}_{vt})^{-1}\hat{\sigma}_e^2$. The summary statistic for v in study t is then the pair $(\hat{\beta}_{vt}, d_{vt})$. One may also estimate $\hat{\beta}_{vt}$ and d_{vt} using a linear mixed model (LMM), which corrects for population structure within the study cohort (Kang *et al.*, 2010; Furlotte and Eskin, 2015).

Because the variance may differ from study to study, we normalize each effect by its standard error to obtain a z-score, where for each variant v, we have

$$z_{vt} = \hat{\beta}_{vt}/\sqrt{d_{vt}} \sim N(\lambda_{vt}, 1) \tag{2.4}$$

where λ_{vt} is the true normalized effect size. One may then use z_{vt} as a test statistic to test whether v is associated with t. Let α be the desired significance level. If $|z_{vt}|$ exceeds some threshold value z_α, or equivalently, $p(z_{vt}) = \Pr(|z| \geq |z_{vt}| \ |H_0) \leq \alpha$, then we conclude v is significantly associated with t.

Because a typical GWAS may test millions of variants, α should be set to account for multiple testing at the variant level. Say 0.05 is the desired significance level for the whole family of tests. A simple way to correct for multiple testing is to apply the Bonferroni correction, which yields $\alpha = 0.05/|V|$. However, due to the presence of linkage disequilibrium

(LD) in the human genome, the Bonferroni correction on the total number of variants is overly conservative. In the GWAS community, $\alpha_{GWAS} = 5 \times 10^{-8}$ is commonly accepted as a significance level that takes into account the number of SNPs and presence of LD in the human genome (Consortium, 2005; Pe'er *et al.*, 2008; McCarthy *et al.*, 2008).

2.2.2 Finding associated variants in at least one of multiple traits using multiple independent GWAS

Suppose we have a variant v and a set of traits $T = \{t_1, ..., t_k\}$, and we are given GWAS effect sizes and variance $(\hat{\beta}_{vt}, d_{vt}^2)$ of v for each trait t in T. To perform multiple independent (MI) GWAS on a set of traits, one simply performs a GWAS as described above for each variant v on each trait to obtain a vector of z-scores across traits $\boldsymbol{z} = (z_{vt_1}, z_{vt_2}, ..., z_{vt_k})^\top$. The MI GWAS test statistic is then $\max_t |z_{vt}|$, or equivalently the smallest GWAS p-value across traits, $\min_t p(z_{vt})$. In MI GWAS, one must correct for two levels of multiple testing - multiple variants and multiple traits. If we assume each trait to be an independent test, then we may apply Bonferroni correction for k traits to α_{GWAS}, yielding multiple testing corrected significance level $\alpha_{MI} = \alpha_{GWAS}/k$. Then v is significant if $\min_t p(z_{vt}) \leq \alpha_{MI}$.

2.2.3 Finding associated variants using CONFIT

CONFIT attempts to find variants $v \in V$ that affect at least one of k traits $t_1, \ldots, t_k$, given summary statistics from a GWAS on each trait. CONFIT assumes each variant v either has zero effect on the trait, or if it has non-zero effect, that its normalized effect size, i.e. its non-centrality parameter (NCP), follows a Fisher polygenic model. We describe whether the

variant has non-zero effect in each of the k traits using a binary vector $\boldsymbol{c} = [c_1 \dots c_k]^\intercal$, where $c_t = 1$ if the variant is active in trait t in that configuration and 0 otherwise.

For convenience, we use a fixed λ for all traits and variants when explaining the test statistic in this section. This fixed λ assumption is very strong. Later we describe how this assumption can be relaxed to allow different NCPs for each variant v. We also assume the z-scores are independent across studies given the activity configuration, but will also relax this assumption in a later section. Let $\boldsymbol{z}_v = [z_{vt_1}, ..., z_{vt_k}]^\top$ Then

$$\boldsymbol{z}_v \sim N(\lambda \boldsymbol{c}, \mathbf{I}) \tag{2.5}$$

Our test statistic at v is a likelihood ratio with multiple alternate models, where model is a different activity configuration. The statistic has the likelihoods of each alternate configuration against $\boldsymbol{c}_0$, weighted by a prior on each configuration $\Pr(\boldsymbol{c})$. Let C denote the set of all possible configurations and $\boldsymbol{c}_0$ denote the null configuration $\boldsymbol{c}_0 = [0 \dots 0]^\intercal$, and C_A denote the set of alternate configurations, $C_A = C \setminus \{\boldsymbol{c}_0\}$. Then

$$F_v = \sum_{\boldsymbol{c} \in C_A} \frac{\mathrm{p}(\boldsymbol{z}|\boldsymbol{c}, \lambda) \Pr(\boldsymbol{c})}{\mathrm{p}(\boldsymbol{z}|\boldsymbol{c}_0) \Pr(\boldsymbol{c}_0)} \tag{2.6}$$

2.2.4 Setting a prior on each activity configuration

Many choices of prior on the configurations are possible. We set an initial prior $\Pr_0(\boldsymbol{c})$ as the fraction of variants which have univariate GWAS p-value less than threshold 10^{-4} in the subset of traits that are active in $\boldsymbol{c}$. We chose 10^-4 as a threshold because we wished to capture shared effects between variants which are not necessarily strong enough

to reach GWAS significance. If $\boldsymbol{c}$ contains only one active trait, we set the final prior $\Pr(\boldsymbol{c})$ by averaging $\Pr_0(\boldsymbol{c}')$ over all configurations $\boldsymbol{c}'$ with a single active trait. Otherwise we set $\Pr(\boldsymbol{c}) = \Pr_0(\boldsymbol{c})$. The reason for this is that the CONFIT model assumes a similar distribution of GWAS z-scores for each trait, but in real life, some traits may tend to have larger effects and others to have smaller effects. We mitigate this by averaging the prior for each trait alone being active. Then traits with large effect sizes will still have high power even with a smaller prior on their configuration, and traits with small effect sizes will now have a power boost with a larger prior. This is the default choice of prior for CONFIT.

2.2.5 Significance testing with F_v

We now describe how to find a p-value and perform significance testing for variant v using F_v.

We find a null distribution for F_v by generating GWAS summary statistics at a variant v under the null hypothesis, by drawing vector of z-scores for each trait $\mathbf{z} \sim N(0, I)$. To generate GWAS summary statistics under the null in the real dataset, one may permute the labels on the set of phenotypes for each trait, such that the correlation between traits is preserved but variant-phenotype correlation is not before performing GWAS, or one could perform GWAS on the real genotypes and simulated phenotypes generated under the null.

The null distribution of F_v also depends on the estimated priors $\{\Pr(\boldsymbol{c}) : \boldsymbol{c} \in C\}$. Say we have estimated priors $\{\Pr(\boldsymbol{c}) : \boldsymbol{c} \in C\}$ from the data. We generate GWAS summary statistics for 5×10^9 variants under the null hypothesis and compute F_v on the null data using the $\{\Pr(\boldsymbol{c}) : \boldsymbol{c} \in C\}$ from the original data. Then we have obtained a null distribution

for F_v. The p-value of F_v, $p(\mathrm{F}_v)$ is the fraction of null variants with test statistic less than F_v. Let p_α be the desired p-value threshold. If $p(\mathrm{F}_v) \leq p_\alpha$, we then conclude variant v is associated with at least one of the k traits.

In a simulated data set containing m independent variants, one may set p_α as the Bonferroni corrected threshold $p_\alpha = 0.05/m$. However, the Bonferroni correction is overly stringent when LD is present between variants, as is the case in real data sets. For the NFBC and UKB data sets, we perform significance testing with F_v at the p-value threshold $p_\alpha = 5 \times 10^{-8}$. This threshold is widely used by the GWAS community to account for multiple testing across the human genome (Pe'er *et al.*, 2008; McCarthy *et al.*, 2008).

2.2.6 Setting a prior on the NCP

We now return to our assumption that NCP $\lambda_{vt} = \lambda$ is fixed for all variants. We instead relax this assumption by allowing each variant to have an NCP drawn from a zero-mean normal distribution with variance σ^2, as in the Fisher polygenic model. Consider a vector of z-scores at the same variant across traits, rather than across variants. Recall our earlier simple formulation, with fixed λ_v for all variants.

$$(\mathbf{z}|\lambda_v, \boldsymbol{c}) \sim N(\lambda_v \boldsymbol{c}, I)$$

This assumption about λ_v is strong and not necessarily realistic. We instead model the NCP for a given variant as a vector, and allow it to differ between traits. Let $\boldsymbol{\lambda}_v = [\lambda_{vt_1}, \ldots, \lambda_{vt_k}]$ be the vector of NCPs across traits for variant v. Supposing a true causal

status $\boldsymbol{c}$, we then put a prior on $\boldsymbol{\lambda}_v$:

$$\boldsymbol{\lambda}_v|\boldsymbol{c} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{k\times k}) \tag{2.7}$$

where $\mathbf{I}_{k\times k}$ is the k-dimensional identity matrix. This prior assumes a Fisher polygenic model on the active traits, where the parameter σ^2 is a fixed value set by the user. In our experiments, we set $\sigma^2 = 25$. However, the performance is not that sensitive to choice of σ^2, as shown in power simulation results for CONFIT with $\sigma^2 = \{4, 10, 36\}$ in Table S2.

2.2.7 Correcting for overlapping individuals across studies

We may also relax the assumption that the estimated effects are independent across traits given the NCPs. This is useful in scenarios where there are overlapping individuals across studies, such as studies where multiple traits are collected from the same individuals. When the cohorts fully overlap between studies (i.e. the k traits are collected from the same individuals), we assume a linear model in each trait

$$\boldsymbol{y}_{t_1} = \beta_{vt_1}\boldsymbol{x}_{vt_1} + \boldsymbol{e}_{t_1}, \ldots \boldsymbol{y}_{t_k} = \beta_{vt_k}\boldsymbol{x}_{vt_k} + \boldsymbol{e}_{t_k} \tag{2.8}$$

where for each individual j, we have $\boldsymbol{y}_j = (\boldsymbol{y}_{t_1,j}, \boldsymbol{y}_{t_1,j})^\top$ following the model

$$\boldsymbol{y}_j = \beta_v\boldsymbol{x}_{v,j} + \boldsymbol{e}_j \tag{2.9}$$

where $\boldsymbol{e}_j \sim N(0, \sigma_e^2 \Sigma_e)$. Σ_e is a k by k covariance matrix representing how the environmental effect on an individual is correlated across traits. Note that under this single-variant linear model,

$$\boldsymbol{\Sigma}_e = \text{Cov}(\boldsymbol{e}_{t_1,j}, ..., \boldsymbol{e}_{t_k,j}) = \text{Cov}(\boldsymbol{y}_{t_1,j}, ..., \boldsymbol{y}_{t_k,j}) \tag{2.10}$$

Let Y be the matrix of phenotype values such that entry y_{ij} is the value of ith trait in the jth individual. The correlation between traits can be modeled as a mix of correlation explained by genetics and correlation explained by shared environment. $\boldsymbol{\Sigma}_e$ should represent correlation explained by the environment. Assume the proportion of covariance explained by genetics is 50%, i.e. each trait in the analysis is 50% heritable. Then $\boldsymbol{\Sigma}_e$ may be estimated as

$$\hat{\boldsymbol{\Sigma}}_e = \frac{1}{2}\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^\top}{n-1} + \mathbf{I}_{k\times k}\right) \tag{2.11}$$

where n is the number of individuals.

If individual level phenotype data is not available, as is often the case with publicly released summary statistics, $\boldsymbol{\Sigma}_e$ may instead be approximated using the correlation between z-scores across traits, assuming that the contribution of any particular variant is small and the heritability is known. Let Z be the matrix of phenotype values such that entry z_{ij} is the value of ith trait in the jth SNP. Then if m is the number of SNPs,

$$\hat{\boldsymbol{\Sigma}}_e = \frac{1}{2}\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^T}{m-1} + \mathbf{I}_{m\times m}\right) \tag{2.12}$$

Under this model with correlated environmental effects for each individual, the distribution of $\boldsymbol{z}_v$ under the null becomes $N(0, \Sigma_e)$ instead of $N(0, I)$, and given a particular alternate

configuration $\boldsymbol{c}$, then $\mathbf{z}|\boldsymbol{c} \sim N(\lambda\boldsymbol{c}, \boldsymbol{\Sigma}_e)$ instead of $N(\lambda\boldsymbol{c}, I)$. We then compute test statistic F_v as in Eq. (2.13) using this distribution for $\mathbf{z}$ to account for correlation due to sharing of individuals between studies.

To generate null CONFIT test statistics to set a significance threshold when studies are correlated, we now draw $\boldsymbol{z} \sim N(0, \Sigma_Z)$, where $\Sigma_Z = \frac{ZZ^T}{m-1}$ is the empirical correlation matrix for the GWAS z-scores. Again assuming that the contribution of any particular variant is small, Σ_Z will capture correlation of z-scores between traits due to the environment and due to variants besides the one being tested.

2.3 Results

2.3.1 Method overview

CONFIT tests whether variant v affects at least one of k traits $t_1, \ldots, t_k$, given summary statistics from a GWAS on each trait. Assume that for each trait, variant v either has an effect on the trait or not, and in each trait where there is an effect, v's non-centrality parameter λ_{vt} (i.e. its standardized effect size or NCP) follows a Fisher polygenic model and is drawn from $\lambda_{vt} \sim N(0, \sigma^2)$. If the variant has non-zero effect on a phenotype, then it is considered "active" in that phenotype. We can then describe a potential activity configuration of a variant in the k traits as a binary vector $\boldsymbol{c} = [c_1 \ldots c_k]^\intercal$, where $c_t = 1$ if it is active in trait t and 0 otherwise. Let C denote the set of all possible configurations, $\boldsymbol{c}_0$ denote the null configuration $\boldsymbol{c}_0 = [0 \ldots 0]^\intercal$, and C_A denote the set of alternate configurations.

The CONFIT test statistic is a sum of the relative likelihoods for each alternate

configuration $\boldsymbol{c}$ against $\boldsymbol{c}_0$, weighted by a prior on each configuration $\Pr(\boldsymbol{c})$:

$$\mathrm{F}_v = \sum_{\boldsymbol{c} \in C_A} \frac{\mathrm{p}(\boldsymbol{z}|\boldsymbol{c}) \Pr(\boldsymbol{c})}{\mathrm{p}(\boldsymbol{z}|\boldsymbol{c}_0) \Pr(\boldsymbol{c}_0)} \tag{2.13}$$

where $\boldsymbol{z} = [z_1, ..., z_k]^\top$ is a vector of standardized GWAS effect sizes for each trait t, $z_t \sim N(\lambda_{vt}, 1)$. The null hypothesis is that v is not active in any trait (corresponding to the null configuration $\boldsymbol{c}_0$), and the alternate hypothesis is that v is active in at least one trait. We estimate the prior on configuration $\boldsymbol{c}$, $\Pr(\boldsymbol{c})$, using GWAS summary statistics for each variant and trait. More details of the method are given in Section 4. We then run CONFIT on simulated datasets to evaluate its performance, and apply it to two real data sets on metabolic traits to find novel variants.

2.3.2 CONFIT increases power when a variant has effect in multiple traits

To measure the power of CONFIT, we generated simulated GWAS summary statistics for k traits as follows. For each variant, we draw a true effect configuration from a multinomial distribution with known probability $\Pr_s(\boldsymbol{c})$ for each configuration $c \in C$, where C is all possible effect configurations. We set $\Pr_s(\boldsymbol{c}) = 0.005$ for each alternate configuration. Then the probability of a variant being active in a given trait is dependent on whether it is active in other traits.

Given the true configuration, for each variant we draw GWAS z-scores with mean zero in traits where there is no effect, and mean $\lambda_s \sim N(0, 25)$ where there is an effect. For each of the following experiments, we generated a panel of 5×10^5 variants. We then run CONFIT

by setting the priors on each configuration from the 5×10^5 variants, then computing the CONFIT test statistic F for each variant. We run this experiment in two and three simulated traits.

The CONFIT test statistic threshold is set using 5×10^9 null simulations for each experiment, and we find no false positives in the these simulations. To demonstrate that the threshold is properly calibrated, we compute the genomic control (GC) factor (Devlin and Roeder, 1999) for CONFIT and for GWAS in each trait in the CONFIT analysis (Tables 2.1,2.2). The GC factor measures how far the median test statistic or p-value deviates from the expected median under the null hypothesis, where larger values indicate more inflation. We find that the GC factor for CONFIT is similar or below the GC factors of the input GWAS. We also show quantile-quantile plots for CONFIT p-values on the North Finland Birth Cohort and UK Biobank data sets in the supplement of Gai and Eskin (2018) in Figure S1.

Table 2.1: Genomic control (GC) factors for the North Finland Birth Cohort (NFBC) data set. We report GC factors for univariate GWAS in each trait and for CONFIT on the glucose (GLU), high-density lipoprotein (HDL), insulin (INS), low-density lipoprotein (LDL), and triglycerides (TG) traits.

Method	GC
GLU	1.000761
HDL	0.998390
INS	1.002076
LDL	0.998764
TG	0.997929
CONFIT	0.841884

From our power simulations, we find that CONFIT loses power compared to MI GWAS when the variant is only active in one trait, but strongly outperforms MI GWAS when the variant is active in more than one trait (Tables 2.3 and 2.4). To understand when CONFIT

Table 2.2: Genomic control (GC) factors for the UK Biobank (UKB) data set. We report GC factors for univariate GWAS in each trait and for CONFIT applied to GWAS summary statistics in four traits.

Method	GC
High cholesterol	1.125458
Cholesterol medication	1.101478
Insulin medication	1.030950
Elevated blood glucose	1.031507
CONFIT	1.106578

has more power over MI GWAS, we plotted the H_0 rejection region for each method on simulated GWAS z-scores in two traits (Figure 2.1a). MI GWAS is slightly more powerful if the GWAS statistic is large in only one trait, but CONFIT is able to detect variants with moderate effects in both traits.

In real datasets, it is possible that some traits will tend have larger or smaller effects than others. To see how CONFIT performs in this case, we also ran simulations where non-zero effects for one trait are drawn from λ_{s1} $N(0, 4)$ and λ_{s1} $N(0, 100)$, and non-zero effects in the remaining traits are drawn $\lambda_s \sim N(0, 25)$. We found that CONFIT still increases power when an effect is present in more than one trait (Table 2.5).

2.3.3 CONFIT increases power in polygenic variants when applied to studies with overlapping cohorts

To model the scenario where each trait is measured in the same cohort, i.e. dependent studies, we simulate summary statistics with correlation Σ_Z between the z-scores across traits, using Σ_Z computed from the Northern Finland Birth Cohort (NFBC) low-density lipoprotein (LDL) and high-density lipoprotein (HDL) traits for simulations in two traits, and from LDL, HDL, and triglycerides (TG) for simulations in three traits. We find that the

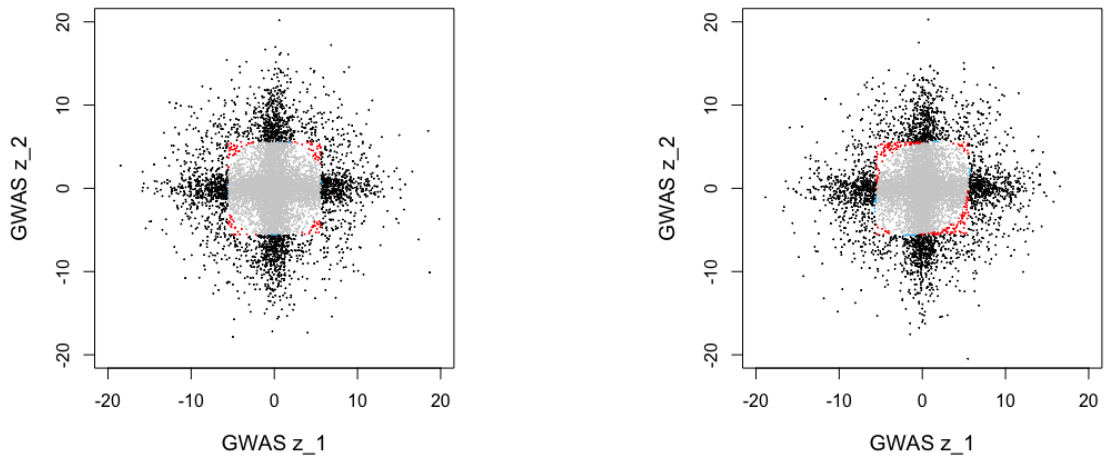


Figure 2.1: Rejection regions for MI GWAS and CONFIT. We ran MI GWAS and CONFIT on simulated GWAS summary statistics in two traits with simulation settings $\lambda_2 \sim N(0, 25)$ for (A) uncorrelated and (B) correlated studies. In each plot, the variants are color coded black if significant by both MI GWAS and CONFIT (i.e. MI GWAS p-value $\leq 2.5 \times 10^{-8}$ and CONFIT p-value $\leq 5 \times 10^{-8}$), red if found significant by CONFIT but not MI GWAS, blue if found significant by MI GWAS and not CONFIT, and grey if not found significant by either method.

Σ_Z estimated from the covariance between individual level phenotypes matches closely with Σ_Z estimated from summary statistics (results not shown). We then run CONFIT with the correction for overlapping individuals described in Section 2.2.7.

Again, we see that CONFIT achieves slightly less power than MI GWAS when the effect is present in one trait, and increased power when the effect is present in more than one trait (Table 2.3 and 2.4). The rejection region for CONFIT is now shifted relative to the rejection region for CONFIT without the overlapping individuals assumption, as shown in Figure 2.1b.

Table 2.3: Power simulation in two traits. Here, the probability of each alternate configuration is set as 0.5%. We draw the true non-centrality parameter (NCP) λ_s for each variant in each trait from a normal distribution for each variant, $\lambda_s \sim \mathrm{N}(0, 25)$, either with or without correlation of effect size between traits. For GWAS in t_1, we only count simulated SNPs which truly have an effect in t_1. We find significant variants using a p-value significance threshold of $5\mathrm{E}{-}08$. For multiple independent (MI) GWAS, we apply the Bonferroni correction to this threshold to account for multiple testing of traits.

	Uncorrelated studies		Correlated studies	
$\lambda_s \sim N(0, 25)$	1 active trait	2 traits	1 trait	2 traits
GWAS in t_1	*0.290*	-	0.291	-
MI GWAS	**0.278**	0.474	**0.283**	0.481
CONFIT	0.272	**0.513**	0.276	**0.540**

Table 2.4: Power simulation in three traits with 0.5% true probability of drawing each alternate configuration. We draw the true NCP λ_s from a normal distribution for each variant, $\lambda_s \sim \mathrm{N}(0, 25)$, either with or without correlation of effect size between traits. For GWAS in t_1, we only count simulated SNPs which truly have an effect in t_1. The power of univariate GWAS in t_1 is in italics. Bolded values indicate multi-trait method with highest power for each simulation.

	Uncorrelated studies			Correlated studies		
$\lambda_s \sim N(0, 25)$	1 active trait	2 traits	3 traits	1 active trait	2 traits	3 traits
GWAS in t_1	*0.283*	-	-	*0.286*	-	-
MI GWAS	**0.274**	0.469	0.607	0.267	0.457	0.602
CONFIT	0.272	**0.504**	**0.681**	**0.285**	**0.518**	**0.697**

2.3.4 CONFIT finds unique loci for metabolic traits in the North Finland Birth Cohort

Next, we applied CONFIT to a real data set, on metabolic traits from the North Finland Birth Cohort (NFBC) dataset (Kang *et al.*, 2010; Sabatti *et al.*, 2008). This dataset contains 331,476 variants and 5,326 individuals, with data collected in ten traits from each individual. These traits include a variety of metabolic traits. We selected the five traits with at least one SNP with a GWAS p-values less than 10^{-4} in two or more traits and ran CONFIT on their summary statistics. These traits were measurements for glucose (GLU), high-density lipoprotein (HDL), insulin level (INS), low-density lipoprotein (LDL), and triglycerides (TG).

Table 2.5: Power simulation in three traits with differing effect size distributions between traits. In the first trait t_1, we draw true effect size $\lambda_{s1} \sim \mathsf{N}(0, 4)$ or $\lambda_s \sim \mathsf{N}(0, 100)$, and in the other two traits, we draw $\lambda_s \sim \mathsf{N}(0, 25)$. The true probability for each alternate configuration is 0.5%. For GWAS in t_1, we only count simulated SNPs which truly have an effect in t_1. The power of univariate GWAS in t_1 is in italics. Bolded values indicate multi-trait method with highest power for each simulation.

$\lambda_{s1} \sim N(0, 4)$	1 active trait	2 traits	3 traits
GWAS in t_1	*0.013*	-	-
MI GWAS	0.182	0.3404	0.474
CONFIT	**0.198**	**0.384**	**0.552**
$\lambda_{s1} \sim N(0, 100)$	1 active trait	2 traits	3 traits
GWAS in t_1	*0.581*	-	-
MI GWAS	**0.366**	0.605	0.768
CONFIT	0.347	**0.627**	**0.832**

Table 2.6: P-values of peak CONFIT SNPs in analysis of five metabolic traits in North Finland Birth Cohort (NFBC) data. Table contains loci found significant by CONFIT or multiple independent (MI) GWAS. The traits used in the analysis are glucose (GLU), high-density lipoprotein (HDL),insulin (INS), low-density lipoprotein (LDL), and triglyceride (TG) levels.

			Univariate GWAS					
Chr	Position	rsID	GLU	HDL	INS	LDL	TG	CONFIT
CONFIT only								
8	19875201	rs10096633	4.5E-01	3.0E-06	4.1E-01	9.3E-01	1.9E-08	**8.0E-10**
16	66570972	rs255049	8.4E-01	1.7E-08	7.3E-01	1.7E-01	1.9E-01	**2.0E-08**
MI GWAS only								
19	11056030	rs11668477	8.3E-01	1.8E-02	1.4E-02	**3.5E-09**	1.7E-02	6.4E-08
Found by both CONFIT and MI GWAS								
1	109620053	rs646776	8.8E-01	1.2E-01	1.0E-01	3.0E-15	7.6E-01	$<$ 2.0E-10
2	21047434	rs6728178	1.6E-01	6.7E-07	8.9E-01	4.8E-08	1.8E-07	$<$ 2.0E-10
2	27584444	rs1260326	2.4E-01	2.6E-01	3.2E-01	2.1E-01	1.9E-10	2.0E-10
2	169471394	rs560887	6.9E-13	8.8E-01	9.9E-01	3.8E-01	6.2E-01	$<$ 2.0E-10
7	44177862	rs2971671	4.4E-09	9.0E-01	2.4E-01	5.9E-01	5.4E-01	8.6E-09
11	92308474	rs3847554	2.4E-10	3.5E-01	1.3E-02	6.2E-01	5.9E-01	8.0E-10
15	56470658	rs1532085	2.3E-01	7.2E-12	5.1E-01	5.6E-01	8.8E-02	$<$ 2.0E-10
16	55550825	rs3764261	4.4E-01	1.0E-32	7.5E-01	2.8E-01	1.2E-01	$<$ 2.0E-10

Note that for MI GWAS with five traits, the significance threshold is 1×10^{-8} for the minimum GWAS p-value out of the five traits.

We used pyLMM (https://github.com/nickFurlotte/pylmm) to obtain GWAS summary statistics on the full NFBC cohort for each trait under a linear mixed model (LMM) as in

(Kang *et al.*, 2010). Our GWAS results are consistent with those reported by a previous GWAS in the NFBC data also using LMMs (Kang *et al.*, 2010). We report the univariate GWAS p-value in each trait as well as the CONFIT p-value in Table 2.6. For MI GWAS in five traits, the significance threshold is 1×10^{-8}.

CONFIT finds two unique loci in the NFBC data compared to MI GWAS. One of these loci (Chr 16, peak SNP rs255049) is significant for HDL under a univariate GWAS threshold, and the other loci (Chr 8, peak SNP rs10096633) has been associated with triglycerides in a larger study from 2010 (Kamatani *et al.*, 2010). CONFIT missed one loci found by MI GWAS only which is GWAS significant for TG only, also shown.

2.3.5 CONFIT outperforms a multivariate linear regression model when applied to multiple traits

Next, we compared the performance of CONFIT against another multi-trait analysis method. Previously, Furlotte et al. applied multivate regression with a linear mixed model (implemented in their software mvLMM) to the NFBC data set using four traits: C-reactive protein (CRP), HDL, LDL, and TG (Furlotte and Eskin, 2015). When running mvLMM to CRP, HDL, LDL, and TG simultaneously, Furlotte et al. found only one significant loci, which contains SNPs rs1811472, rs2794520, rs2592887, and rs12093699.

We applied CONFIT to the NFBC data set in these same four traits, again using GWAS summary statistics generated by pyLMM. CONFIT in fact finds this loci, as well as nine other loci (Table 2.7) which were all reported in the univariate LMM analysis performed by (Kang *et al.*, 2010). CONFIT discovers the same loci in these four traits as in the analysis

on GLU, HDL, INS, LDL and TG, with the exception of a GLU-specific locus. It also finds a loci (Chr 19, rs11668477) that it missed in the five trait analysis. Although CONFIT can discover SNPs with effects present in only a subset of traits in the analysis, the specific traits chosen will affect its performance.

Table 2.7: P-values of peak CONFIT SNPs in analysis of four metabolic traits in North FInland Birth Cohort data set. Table contains peak CONFIT SNPs for loci found significant by CONFIT or MI GWAS. Italics indicates the only loci found significant by (Furlotte and Eskin, 2015) in their joint analysis of all four traits.

			Univariate GWAS				
Chr	Position	rsID	CRP	HDL	LDL	TG	CONFIT
CONFIT only							
8	19875201	rs10096633	3.9E-01	3.0E-06	9.3E-01	1.9E-08	4.0E-09
16	66570972	rs255049	7.8E-01	1.7E-08	1.7E-01	1.9E-01	4.2E-08
Found by both CONFIT and MI GWAS							
1	109620053	rs646776	1.4E-01	1.2E-01	3.0E-15	7.6E-01	$<$ 2.0E-10
1	157908973	*rs1811472*	1.2E-15	4.8E-02	6.1E-01	8.7E-01	$<$ 2.0E-10
2	21047434	rs6728178	5.3E-02	6.7E-07	4.8E-08	1.8E-07	$<$ 2.0E-10
2	27584444	rs1260326	5.1E-02	2.6E-01	2.1E-01	1.9E-10	2.4E-09
12	119873345	rs2650000	2.2E-12	2.8E-01	6.8E-01	6.0E-01	$<$ 2.0E-10
15	56470658	rs1532085	7.1E-01	7.2E-12	5.6E-01	8.8E-02	$<$ 2.0E-10
16	55550825	rs3764261	3.2E-01	1.0E-32	2.8E-01	1.2E-01	$<$ 2.0E-10
19	11056030	rs11668477	8.7E-01	1.8E-02	3.5E-09	1.7E-02	3.4E-08

2.3.6 CONFIT finds unique loci in the UK Biobank dataset

We also applied CONFIT to UK Biobank summary statistics publicly released by Neale lab. We selected four traits related to the metabolic traits we used in the NFBC data. These are: self-reported high cholesterol (phenotype code 20002_1473), use of cholesterol lowering medication (phenotype code 6177_1), use of insulin medication (phenotype code 6177_3), and diagnosis of elevated blood glucose level (phenotype code R73, ICD10 R73). CONFIT finds 6

unique loci (2.8), MI GWAS finds 44 unique loci (shown in Supplementary materials of (Gai and Eskin, 2018)), and 304 loci are found by both methods (not shown). The loci found by CONFIT are all close to GWAS significance in both the self-reported high cholesterol and use of cholesterol medication phenotypes, whereas the loci it fails to discover are mostly borderline GWAS significant in a single trait (Table S1).

Table 2.8: P-values of peak SNPs in analysis of four metabolic traits in UK Biobank data set. Table contains peak SNPs found significant by CONFIT (CONFIT p-value $\leq$ 5E-08) only. SNPs found significant by MI GWAS only are shown in the Supplementary materials of (Gai and Eskin, 2018).

			Univariate GWAS				
Chr	Position	rsID	High cholesterol	Cholesterol medication	Insulin medication	Elevated blood glucose	CONFIT
CONFIT only							
3	135925191	rs1154988	5.2E-08	9.8E-07	7.2E-01	2.3E-01	5.6E-09
7	73020301	rs799157	3.2E-08	3.1E-05	5.4E-01	7.3E-01	3.9E-08
7	150690176	rs3918226	3.0E-08	3.0E-07	3.9E-01	1.9E-01	1.0E-09
10	94772638	rs10748588	2.3E-07	1.5E-06	8.7E-01	3.1E-01	3.0E-08
11	126225876	rs112771035	5.9E-07	4.2E-06	3.6E-02	8.0E-01	4.8E-08
20	17844492	rs2618567	1.6E-08	6.0E-07	3.9E-01	2.4E-01	1.0E-09

2.4 Discussion

Here we present CONFIT, a method for detecting associated variants from independent GWAS in multiple traits using summary statistics. We demonstrate our method in simulated data on two and three traits, and on real data up to four traits, though this framework may be applied to larger numbers of traits. CONFIT controls the false positive rate and increases power relative to multiple independent (MI) GWAS when the variant is active in multiple traits in the analysis. When the variant is only active in one trait, CONFIT is less powerful than MI GWAS, which is the standard method for analyzing independent traits, so CONFIT

does not discover exactly the same SNPs as GWAS. We discover unique loci when applying CONFIT to summary statistics from the NFBC and UK Biobank data sets.

A related problem exists in the field of eQTL studies, which often collect gene expression data from individuals in multiple tissues. In this case, the phenotypes are a given gene's expression levels in each tissue, and the problem is to find variants associated with the gene's expression in at least one tissue. Several approaches have successfully increased power in these multi-tissue eQTL data sets. Examples include MetaTissue (Sul *et al.*, 2013), RECOV (Duong *et al.*, 2017), and eQTL-bma (Flutre *et al.*, 2013). MetaTissue uses random effects meta-analysis to combine data from different tissues. RECOV explicitly models correlation between studies using a covariance matrix. eQTL-bma uses configurations to allow heterogeneity and performs Bayesian model averaging using each potential configuration as a model. We note the similarity of our test statistic to that of eQTL-bma, which was developed by Flutre et al. specifically for multi-tissue eQTL context (Flutre *et al.*, 2013). A variant is an eQTL if it is associated with the expression of any gene in any tissue, which is quite likely when there is a large number of tissues. For this reason, methods developed for multi-tissue eQTL studies differ from those for traditional GWAS in that eQTL studies typically do not assume a sparse model. In contrast, the majority of variants are believed to have no effect on the majority of disease traits. Hence it is not obvious whether multi-phenotype analysis methods for eQTL studies are also applicable to GWAS. Our results suggest they may be applicable.

The CONFIT framework is general and there are many options for setting the priors on each configuration. Here we used a relatively simple method to estimate the priors by counting the number of SNPs with GWAS summary statistics that match each configuration.

One alternative is to formulate this as an optimization problem and select priors that explicitly maximize power, with some form of regularization to avoid overfitting. Another possibility is to use external information about the variants to set the prior. This has been done previously in eQTL data, where variants in regulatory regions receive a stronger prior for association (Duong *et al.*, 2016).

The count-based prior used here has the disadvantage of not scaling well as the number of traits grows, since as the number of possible configurations grows exponentially, the probability of observing any particular configuration decreases sharply. From a methods viewpoint, count-based methods for setting the prior on each configuration become less and less useful with larger numbers of traits, as the probability of observing any particular configuration amongst the GWAS statistics decreases with the number of traits. From a computational viewpoint, the runtime of CONFIT grows exponentially. For these reasons, we do not recommend running CONFIT on more than 10 traits. If the user has a large set of candidate traits, they may narrow down which traits to include in the analysis by choosing sets of traits with overlapping GWAS significant SNPs. One may use the Jacquard index to measure overlap between traits while also accounting for the fact where one trait may simply have more significant SNPs than other traits.

It is common for GWAS datasets to share individuals between studies. For example, a study may collect both LDL and triglyceride levels from each individual, or controls may be shared across multiple case-control studies. CONFIT handles the cases where the studies use the same cohort by approximating the correlation between traits due to sharing of individuals as proportional to correlation between traits or association statistics. This assumes the effect and residuals are approximately independent, and that any individual SNP or LD

block has small effect on the phenotype. In this paper, we assume heritability of 50% when estimating this correlation, but a more sophisticated approach would be to use trait-specific heritability estimates. There are also many other methods to address the issue of overlapping individuals. For example, MetaTissue uses linear mixed models (LMMs) to model effects in multiple studies with shared individuals (Sul *et al.*, 2013). Although their method was designed for multi-tissue eQTL studies, a similar LMM approach could be applied to combine GWAS. This approach has the advantage of estimating the proportion of the phenotype that can be attributed to sharing of individuals, and applies even if there is only partial overlap between studies. However, it requires individual level data and is relatively computationally expensive.

Several methods for analyzing multiple traits require individual level genotype and phenotype data, such as multivariate regression. Several methods, such as GEMMA-mvLMM, mvLMM, and GAMMA, extend this to use linear mixed models, which allow for correction of population structure and other covariates (Zhou and Stephens, 2014; Furlotte and Eskin, 2015; Joo *et al.*, 2016). As with traditional meta-analysis, multivariate regression is not suitable for combining data on arbitary traits and may achieve suboptimal power for detecting variants that only affect one of the traits tested, or in the case where the variant only affects one trait, which indirectly affects another (Stephens, 2013). Such methods are typically applied to sets of traits that are already believed to share an underlying genetic basis (Furlotte and Eskin, 2015). Thus there is a need for flexible approaches to association testing when the traits only partially share a genetic basis and the study cohorts are not independent between traits.

Chapter 3

Evaluating models for variant effects across multiple traits using using summary statistics

3.1 Introduction

Polygenic scores (PGS) are commonly used for the study of genetic architecture, e.g. (Purcell *et al.*, 2009; Jones *et al.*, 2018; Del-Aguila *et al.*, 2018) and have also emerged as a promising tool for disease risk assessment, e.g. (Shieh *et al.*, 2016; Logue *et al.*, 2018). PGS combine genetic information from many variants, typically single-nucleotide polymorphisms (SNPs), allowing them to provide meaningful estimates of genetic liability even when any single variant has a small contribution to the disease or trait of interest.

The effect sizes used to compute PGS are typically estimated using genome-wide association studies (GWAS), which fit a linear model for each SNP in the study on the

trait. However, small effects are difficult to estimate accurately using traditional GWAS, sometimes requiring hundreds of thousands or millions of samples, in phenotypes that may be difficult to collect.

One approach to improve GWAS estimates is to combine information from multiple related traits. Many pairs of traits exhibit genetic correlation, i.e. their effect sizes are correlated, and exhibit significant genetic correlation even in the absence of any significantly associated loci (Bulik-Sullivan *et al.*, 2015a; Shi *et al.*, 2017). Several existing methods leverage this genetic correlation in multiple traits to estimate variant effects from summary statistics (Hu *et al.*, 2017; Maier *et al.*, 2018; Turley *et al.*, 2018; Qi and Chatterjee, 2017). In particular, the method Multi-Trait Analysis of GWAS (MTAG) (Turley *et al.*, 2018) requires only summary statistics and has already been applied in a variety of settings (Lam *et al.*, 2017; Grove *et al.*, 2019).

MTAG assumes the genetic correlation across traits is identical across the genome, and that all SNPs have an effect in all traits. In cases where the assumption is violated, e.g. if a variant only has an effect in a subset of the traits, MTAG may produce biased estimates of effect size (Turley *et al.*, 2018). It has been shown that the genetic correlation between traits can vary from region to region (Shi *et al.*, 2017) and the distribution of SNP effect sizes varies with minor allele frequency (MAF) and degree of linkage disequilibrium (LD) (Evans *et al.*, 2018; Pazokitoroudi *et al.*, 2020). Based on these observations, we decided to evaluate models for effect sizes across traits which allow for sparsity or different patterns of correlation across the genome, such as Gaussian mixture models (GMMs) and models stratified by MAF and degree of LD.

Here we compare several such models against GWAS, which is on a single trait at a

time, and MTAG, which assumes identical covariance across the genome. We evaluate these methods by using them to compute PGS on anthropometric and mental health traits from the UK Biobank (UKB) (Bycroft *et al.*, 2018). We find that GWAS estimates sometimes produce more accurate PGS than the multi-trait methods even when the genetic correlation between traits is strong, though MTAG and the MAF and LD stratified model often outperformed multiple variance component models. We hope our findings will help inform the use cases for multi-trait approaches to PGS and future work on estimating variant effects.

3.2 Methods

3.2.1 Association testing and polygenic model for a single trait

We describe how to perform a genome-wide association study (GWAS) at a SNP j on a single trait t, using data from N individuals. Suppose we have a vector of standardized genotypes $\boldsymbol{x}$ at SNP j at each individual, and a vector of standardized phenotypes $\boldsymbol{y}_t$ for each individual in trait t. Then we may estimate the scalar effect β_j of SNP j on trait t using linear regression:

$$\hat{\beta}_{jt} = \frac{1}{N}\mathbf{x}^\top \boldsymbol{y} \sim N\left(\beta_{jt}, \frac{1}{N}\sigma^2_{e_j}\right) \tag{3.1}$$

where $\sigma^2_{e_j}$ is the variance of contributions to the phenotype from factors besides SNP j. The variance $\sigma^2_{e_j}$ may be estimated as $\hat{\sigma}^2_{e_j} = \frac{1}{N}(\boldsymbol{y} - \hat{\beta_{jt}}\boldsymbol{x}_j)^\top(\boldsymbol{y}_t - \hat{\beta_{jt}}\boldsymbol{x}_j)$.

GWAS assumes a linear model where only one SNP has non-zero effect, so the effects of any SNPs in LD with SNP j are also captured in β_j and subsequently in $\hat{\beta_{jt}}$. For this reason, β_{jt} is used here to refer to the marginal effect of SNP j. To compute GWAS estimates for

this paper, we used the software PLINK (version 1.9) (Purcell *et al.*, 2007).

Next, we describe the additive polygenic model used to compute a polygenic score (PGS). In this model, an individual's phenotype is simply the weighted sum of their standardized genotypes at a set of SNPs, say a set of M SNPs. Recall that each β_{jt} has been estimated on standardized phenotypes. Then the (standardized) phenotype for individual i in trait t is given by

$$y_{it} = \sum_{j=1}^{M} \beta_{jt} x_{ij} + e_i \tag{3.2}$$

where $e_i \sim N(0, \sigma_e^2)$ is environmental effects. It is also assumed that $\sum_{j=1}^{M} \beta_j \sim N(0, \sigma_g^2)$, with narrow heritability $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$ corresponding to the fraction of phenotypic variance explained by additive SNP affects.

To estimate the PGS for an individual i from a set of M SNPs, we compute

$$\hat{y}_{it} = \sum_{j=1}^{M} \hat{\beta}_{jt} x_{ij}. \tag{3.3}$$

There are several options for choosing the set of SNPs to use in the PGS. For example, one may use all genotyped SNPs, SNPs that are GWAS-significant, or some LD-pruned subset of SNPs. If using non-independent SNPs as predictors for the PGS, we must apply a correction for LD so that the summation is over the non-marginal SNP effects. Otherwise, SNP effects will effectively be counted multiple times in the summation. Several methods exist for performing this correction when computing the PGS, such as LDpred (Vilhjálmsson *et al.*, 2015) which was used for PGS calculation by Turley *et al.* (2018).

3.2.2 Multivariate normal model for effects across multiple traits

Suppose that we have GWAS summary statistics from K traits at SNP j. This consists of GWAS estimated effect sizes $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jK})$, as well as the sample variance of these estimates.

Also suppose that the true marginal SNP effects across traits $\boldsymbol{\beta_j} = (\beta_{j1}, \ldots, \beta_{jK})^\top$ are drawn from a multivariate normal (MVN), such that

$$\beta_j \sim N(\mathbf{0}, \boldsymbol{\Omega}) \tag{3.4}$$

where $\mathbf{0}$ is a vector of all zeros of appropriate dimension and $\boldsymbol{\Omega}$ is the genetic covariance matrix, such that entry ω_{ij} is proportional to the genetic correlation between traits i and j. Note that $\beta_{\mathbf{j}}$ is a vector of marginal SNP effects, that is, it includes the effects of other SNPs in LD with the SNP of interest.

Suppose we then have K studies, one for each trait. Given the true effects for SNP j, the GWAS linear estimator will come from the following distribution:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} | \boldsymbol{\beta_j} \sim N(\boldsymbol{\beta_j}, \boldsymbol{\Sigma}_j) \tag{3.5}$$

Where $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ is the vector of estimated effects across studies for SNP j, and $\boldsymbol{\Sigma}_j$ is the covariance matrix corresponding to the estimation error, i.e. the variance-covariance matrix for estimation errors for SNP j. For each pair of traits (t, s), the entry $\Sigma_{j,t,s}$ is proportional to the environmental correlation across studies for this pair of trait, which may be non-zero if there is sample overlap across studies.

In practice, $\mathbf{\Omega}$ and $\mathbf{\Sigma}_j$ are not known. We estimate $\mathbf{\Omega}$ using the methods of moments, as described in Turley *et al.* (2018):

$$\hat{\mathbf{\Omega}} = \frac{1}{M}\sum_{j=1}^{M}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}{}' - \hat{\mathbf{\Sigma}}_j\right) \tag{3.6}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ is the vector of GWAS estimates for SNP j.

We also estimate the entries of $\mathbf{\Sigma}_j$ as in Turley *et al.* (2018). Suppose $N_{t,j}$ and $N_{s,j}$ are the study sizes at SNP j corresponding to trait t and trait s. Then we estimate the corresponding entries in $\mathbf{\Sigma}_j$ as:

$$\hat{\Sigma}_{j,t,t} = \hat{\sigma}^2_{e_t}/N_{t,j}\hat{\Sigma}_{j,t,s} = \hat{\sigma}_{e_t}\hat{\sigma}_{e_s}/\sqrt{N_{t,j}N_{s,j}} \tag{3.7}$$

where $\sigma^2_{e_t}$ and $\sigma_{e_t}\sigma_{e_s}$ are variance and covariance due to sample overlap between the two studies. These values can be estimated from the LD score regression intercepts Bulik-Sullivan *et al.* (2015a).

3.2.3 MTAG estimator for effect size

Here we describe the multi-trait linear estimator derived by Turley *et al.* for estimating effects from GWAS summary statistics, assuming all effects across the genome share the same variance-covariance matrix $\mathbf{\Omega}$ (Turley *et al.*, 2018).

Suppose we have GWAS summary statistics for a trait of interest (the "primary" trait) and additional related traits (the "auxillary" traits), as well as covariance matrices $\mathbf{\Omega}$ and $\mathbf{\Sigma}$. For purposes of the derivation, we assume that the true values of $\mathbf{\Omega}$ and $\mathbf{\Sigma}$ are known.

Denote the GWAS effect sizes from K traits as $\hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jK}$. (Turley *et al.*, 2018) derive an estimator b_{MTAG} for the effect in the primary trait β_1:

$$b_{MTAG} = \frac{\left(\frac{\omega_1}{\omega_{11}}\right)^\top \left(\mathbf{\Omega} - \frac{1}{\omega_{11}}\omega_1\omega_1^\top + \mathbf{\Sigma}_j\right)^{-1} \hat{\boldsymbol{\beta}}}{\left(\frac{\omega_1}{\omega_{11}}\right)^\top \left(\mathbf{\Omega} - \frac{1}{\omega_{11}}\omega_1\omega_1^\top + \mathbf{\Sigma}_j\right)^{-1} \left(\frac{\omega_1}{\omega_{11}}\right)} \tag{3.8}$$

Note that MTAG estimates the effects of all traits in the analysis jointly, but for consistency with other methods we test, we use the primary/auxiliary trait notation.

3.2.4 Two component mixture models for effects across traits

Component models allow us to relax the assumption that SNP effects are drawn from the same distribution across the genome. In this section, we describe four models where the SNP effects across traits are drawn from a Gaussian mixture model with two components.

Let $\beta_j = (\beta_{j1}, \ldots, \beta_{jK})^\top$ denote the true marginal effects of SNP j in traits $1, \ldots, K$. Let π_0 and π_1 be the mixing weights of each component, such that $\pi_0 = 1 - \pi_1$, and let γ_j be a latent variable for which component SNP j was drawn from. Say the trait of interest t_1 is the first entry in β, which we will refer to the primary trait, and other traits as auxiliary traits. The generative model for the true marginal effects at a SNP j is as follows.

$$\gamma_j \sim \text{Bernoulli}(\pi) \tag{3.9}$$

$$\beta_j | \gamma_j = 0 \sim N\left(\mathbf{0}, \mathbf{\Omega}_0\right) \tag{3.10}$$

$$\beta_j | \gamma_j = 1 \sim N\left(\mathbf{0}, \mathbf{\Omega}_1\right) \tag{3.11}$$

where $\mathbf{\Omega}_0$ is a K by K matrix chosen from one of the options described below, and $\mathbf{\Omega}_1$ is a non-sparse covariance matrix.

We set $\mathbf{\Omega}_1$ to correspond to the classic polygenic model with full genetic correlation, estimated using Eq. 3.6, while $\mathbf{\Omega}_0$ contains a subset of entries of $\mathbf{\Omega}_1$. We test four options for $\mathbf{\Omega}_0$ each corresponding to a different possible GMM, described below.

Denote the entries of $\mathbf{\Omega}_1$ as:

$$\mathbf{\Omega}_1 = \begin{bmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 & \cdots & \rho_{1k}\tau_1\tau_k \\ \rho_{12}\tau_1\tau_2 & \tau_2^2 & \cdots & \rho_{2k}\tau_2\tau_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1k}\tau_1\tau_k & \rho_{2k}\tau_2\tau_k & \cdots & \tau_k^2 \end{bmatrix}$$

We test the following four choices of $\mathbf{\Omega}_0$, each reflecting different assumptions about the relationships between traits in the analysis.

– If we assume the primary trait has an effect and the other traits do not (we will refer to the corresponding model as GMMa):

$$\mathbf{\Omega}_{0a} = \begin{bmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \epsilon & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \epsilon \end{bmatrix}$$

where ϵ is an arbitrarily small value.

– If we assume the primary trait is sometimes sparse, but correlation amongst the auxillary traits is the same for all SNPs (GMMb):

$$\boldsymbol{\Omega}_{0b} = \begin{bmatrix} \epsilon & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \cdots & \rho_{2k}\tau_2\tau_k \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \rho_{2k}\tau_2\tau_k & \cdots & \tau_k^2 \end{bmatrix}$$

where again ϵ is an arbitrarily small value.

– If we assume there is no sparsity in effect sizes, but the primary trait is sometimes uncorrelated with the auxillary traits (GMMc):

$$\boldsymbol{\Omega}_{0c} = \begin{bmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \cdots & \rho_{2k}\tau_2\tau_k \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \rho_{2k}\tau_2\tau_k & \cdots & \tau_k^2 \end{bmatrix}$$

– If we assume there is no sparsity in effect sizes, but effect sizes are sometimes completely

independent in all traits (GMMd):

$$\mathbf{\Omega}_{0d} = \begin{bmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tau_k^2 \end{bmatrix}$$

Given $\mathbf{\Omega}_0, \mathbf{\Omega}_1$, the mixing parameters π_0, π_1 are estimated by expectation-maximization (EM). In the E step, we compute membership assignment for each SNP based on our current estimates for the mixture weights. Use the notation $N(\hat{\boldsymbol{\beta}}_j|\mathbf{0}, \mathbf{A})$ to denote the density at $\hat{\boldsymbol{\beta}}_j$ (the observed GWAS estimates for SNP j) of a centered multivariate normal distribution with covariance matrix $\mathbf{A}$. Then the E step is:

$$p(\gamma_j = 0|\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \mathbf{\Omega}_0, \mathbf{\Sigma}_j) \leftarrow \frac{N(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}|\mathbf{0}, \mathbf{\Omega}_0 + \mathbf{\Sigma}_j)\pi_0}{\sum_{c=0}^{1} N(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}|\mathbf{0}, \mathbf{\Omega}_c + \mathbf{\Sigma}_j)\pi_c} \tag{3.12}$$

for $j = 1, \ldots, M$.

In the M step, we total the fraction of SNPs assigned to each component to get the mixture weights. The M step becomes

$$\pi_0 \leftarrow \frac{\sum_{j=1}^{M} p(\gamma_j = 0|\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \mathbf{\Omega}_0, \mathbf{\Sigma}_j)}{M} \tag{3.13}$$

for $j = 1, \ldots, M$.

We alternate between the E and M steps until convergence. These EM updates assume

that each SNP is an independent sample from the mixture distribution, so they should be computed over an LD-pruned subset of all available SNPs. In practice, we use estimates of $\mathbf{\Omega}_0$ and $\mathbf{\Omega}_1$ for the EM process.

3.2.5 Stratifying SNPs by LD score and MAF

The four GMMs in the previous section base both covariance components on the GWAS estimated effect sizes and covariance intercepts across all available SNPs. As an alternative approach, we also tested a model where each component corresponds to a different subset of SNPs, stratified by minor allele frequency (MAF) and LD score. Rare SNPs tend to have a different effect size distribution than common SNPs, and SNP marginal effects include the effects of any SNPs in strong LD.

We tested a model with SNPs stratified into four bins. In this model, we split SNPs into top 50% or bottom 50% by LD score, and further split into MAF < 0.1 or MAF ≥ 0.1. We obtained estimates of covariance due to sample overlap for each bin using LDSC (Bulik-Sullivan *et al.*, 2015b,a). We estimated a genetic correlation matrix for each bin as in Eq. 3.6, except now only using the subset SNPs in a bin rather than using all SNPs. Then we obtained the MTAG effect estimates (Eq. 3.8) for each SNP using the genetic covariance and sample overlap estimated for that particular SNP's bin.

3.2.6 SNP filtering and computing polygenic scores

In selecting which SNPs to include in our analysis, we applied the SNP filter used by Turley *et al.* (2018) for SNP discovery and effect size estimates. In addition to standard QC, this

filter removes SNPs with MAF $< 1\%$ or much lower sample size than other SNPs for that trait, and SNPs that are outliers with respect to effect size, including an inversion region in Chromosome 8 strongly associated with neuroticism. This filter was applied before running GWAS, and thus before estimating genetic correlation and covariance due to sample overlap from the GWAS effects.

In the PGS formula in Eq. 3.2, it was assumed the SNPs to compute the PGS are independent. However in practice, there is widespread correlation between SNPs in the genome, i.e. linkage disequilibrium (LD). To account for LD, one may either select a subset of approximately independent SNPs using LD pruning and thresholding by GWAS p-value, or adjust the marginal SNP effects for LD before computing PGS. We used LDpred (version 1.06) to adjust estimates of marginal SNP effects for LD before computing the PGS (Vilhjálmsson *et al.*, 2015), as in Turley *et al.* (2018). LDpred takes in an LD reference panel, an LD radius, and assumed fraction of casual SNPs. We ran LDpred with a random sample of 5,000 individuals who were excluded from the PGS cohort as the LD reference panel. The LD radius was set to 150 and the fraction of causal SNPs was assumed to be 1 (the infinitesimal model setting). We applied LDpred to effect estimates from GWAS, MTAG, and our five additional models to compute PGS using each method.

3.3 Results

3.3.1 Overview

We applied GWAS, MTAG, and our five additional multi-trait models (four GMMs using the genetic correlation matrix estimated from all SNPs, and one based on stratifying SNPs by MAF and LD) to estimate SNP effect sizes and compared the predictive power of the resulting PGS from each method.

For the following experiments, we only used unrelated white British individuals in the UK Biobank. We randomly subsampled 200k individuals from the UK Biobank to use as a GWAS cohort and 10k individuals to use as a PGS cohort so that individuals used to obtain effect size estimates were not used to measure predictive power. Note that we did not take missing phenotypes into account when sampling, so the final GWAS sizes are proportionate to the original sample sizes. We use LDpred (Vilhjálmsson *et al.*, 2015) to estimate PGS from effect sizes, as in Turley *et al.* (2018). We then computed the Pearson correlation between the PGS from each method and the true phenotypes of the PGS cohort.

3.3.2 PGS on anthropometric and psychiatric traits from the UKB

We applied these methods to four sets of anthropometric traits in the UK Biobank chosen to represent different scenarios for multiple traits. All traits used in these experiments had on the order of 200k individuals with non-missing phenotype values in the GWAS cohort, except for manual pulse and systolic blood pressure, which had on the order of 20k individuals.

GWAS sample sizes are shown in Table. 3.1. We used age, sex, and the first 20 genetic prinicipal components (PCs) as the covariates when computing the GWAS estimates.

The sets were (1) arm fat percentage in left arm, trunk fat percentage and waist circumference; (2) automated pulse measurement and pulse rate (during blood-pressure measurement); (3) automated pulse measurement and standing height, (4) automated pulse measurement, standing height, and seated height. We will refer to the "pulse rate (during blood-pressure measurement)" trait as manual pulse for convenience. These sets were chosen to represent these scenarios respectively: (1) strong genetic correlation between all traits and large sample size for all traits, (2), strong genetic correlation where one trait has a much smaller sample size than the other, (3) weak genetic correlation between two traits both with large sample size, and (4) a combination of strong and weak genetic correlations with varying sample sizes. Each trait in each set was used as the primary trait in turn. The genetic correlations between traits are shown in Table. 3.2.

We then computed the Pearson correlation between the PGS from each method and the true phenotypes of the PGS cohort. We report incremental R^2, which is the proportion increase in R^2 between the PGS estimated GWAS or one of the multi-trait methods, compared to prediction using linear model with covariates only (Figs. 3.1a-3.1d).

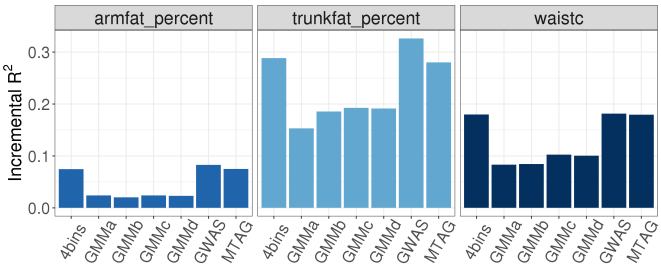
We found that GMMs A-D typically performed similarly to each other, with GMMb (a model where effects in the primary trait may be sparse) outperforming other GMM methods on a few sets.

MTAG and the stratified model's performance varied across settings. MTAG consistently outperformed the GMMs and stratified model, and that its performance relative to GWAS was strongest for predicting manual pulse using automated pulse (Fig. 3.1b), though it did

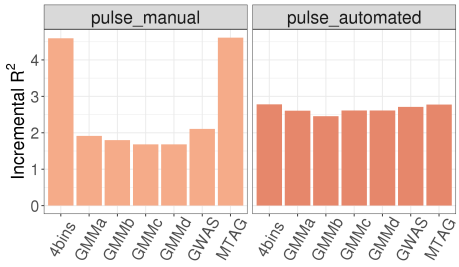
underperform GWAS for both traits in the pulse and height set. We suspect that MTAG is mainly useful in situations where GWAS in the trait of interest is underpowered, but otherwise performs comparably to GWAS, or worse if the traits are weakly correlated.

Trait	N
armfat_percent	196,512
height	199,577
height_seated	199,570
pulse_automated	189,901
pulse_manual	18,803
sysbp	18,803
trunkfat_percent	196,465
waistc	199,658
DEP	180,263
NEU	162,838

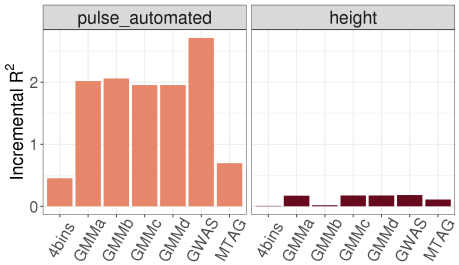
Table 3.1: Sample size for GWAS on anthropometric and psychiatric traits in UKB. We subsampled 200k individuals to form a GWAS cohort, then performed a GWAS for each trait using the non-missing individuals for that trait. The GWAS cohort was excluded when estimating polygenic risk scores.



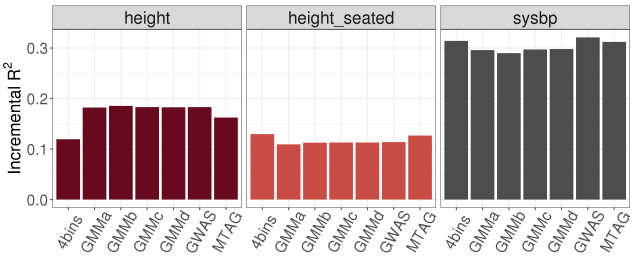
(a) Three traits with strong genetic correlation and similarly large sample sizes.



(b) Two traits with strong genetic correlation, where pulse_manual has smaller sample size.



(c) Two traits with weak genetic correlation.



(d) Three traits where only two traits have strong genetic correlation with each other.

Figure 3.1: Predictive power of mixture models, GWAS, and MTAG on anthropometric traits. Polygenic scores (PGS) were computed using a stratified model, one of four GMMs, GWAS, or MTAG effect size estimates. For the multi-trait methods, each figure is labeled with the primary trait, with the other traits in the same set used as auxiliary traits. The sets of traits used were (a) arm fat percent, trunk fat percent, and waist circumference; (b) automated pulse and manual pulse; (c) automated pulse and standing height, (d) automated pulse, standing height, and seated height. Incremental R^2 is proportion increase in R^2 between the PGS and observed phenotypes, compared to PGS estimated using linear model with covariates only.

In addition to the four sets of anthropometric traits, we also analyzed a set of two psychiatric traits from the UKB, depression and neuroticism. As before, we split the samples into a GWAS cohort and PGS cohort and conducted GWAS, using the same procedure as for the anthropometric traits. We found that GWAS performed comparably to the the GMMs on DEP and outperformed other methods on NEU (Figure 3.2).

	DEP	NEU
DEP	1(0)	0.82(0.03)
NEU	0.82(0.03)	1(0)

Table 3.3: Genetic correlation for neuroticism and depression, estimated using LDSC software package on the same individuals and filtered SNPs used for GWAS.

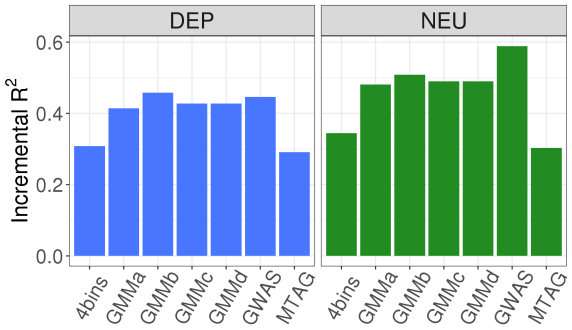


Figure 3.2: Predictive power of mixture models, GWAS, and MTAG on a set of two psychiatric traits, depression and neuroticism. Polygenic scores (PGS) were computed using effect estimates from one of four GMMs, a stratified model, GWAS, or MTAG.

3.3.3 PGS improvement from multitrait prediction varies with study sizes

To test whether multi-trait methods did have more of an impact on PGS prediction when the trait of interest is underpowered, we repeated our earlier experiments on the highly correlated arm fat, trunk fat, and body weight traits, but now downsampling the traits to yield varying study sizes. Consistent with our earlier results, we find that MTAG and the stratified model outperform GWAS when the primary trait has a small study and the auxiliary traits are large, but underperform GWAS when the primary trait has a large study (Fig. 3.3). Interestingly, MTAG and the stratified model also underperformed GWAS when using weight as the primary trait when all three studies are small.

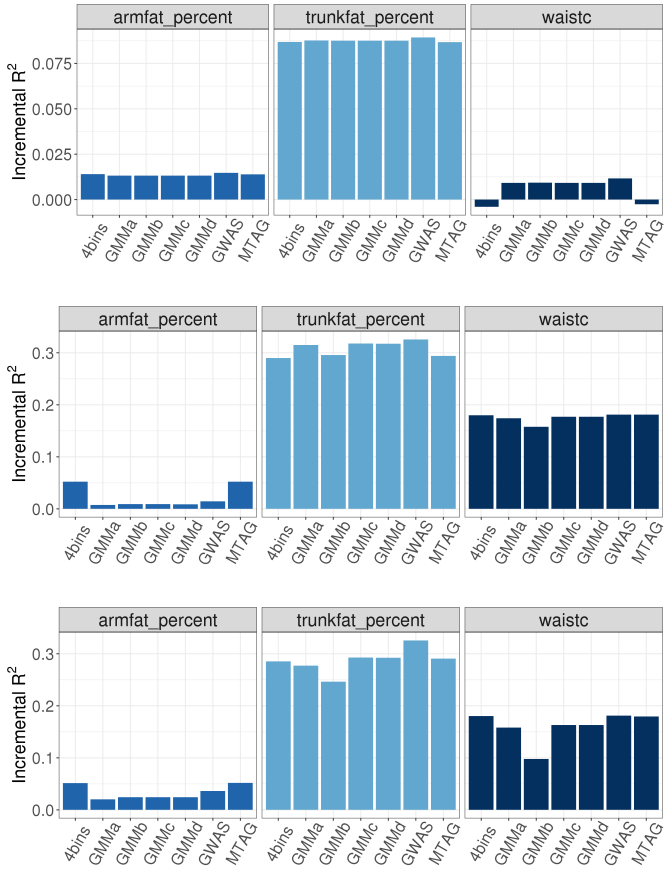


Figure 3.3: Effect of study size on incremental R^2 between PGS and observed phenotypes, which measures the increase in R^2 when using genetics compared to a prediction based on covariates only. The same three traits were used as a set for three experiments where each trait's GWAS effect size estimates were re-estimated from cohorts of varying size before being passed to the multi-trait methods as input. Traits used were arm fat percent (left), trunk fat percent (middle), and waist circumference (right). (a) All traits with GWAS of 20k samples. (b) Arm fat percent with 20k samples, and the others with 200k samples each. (c) Arm fat percent with 50k samples, and others with 200k samples each.

	armfat_percent	height	height_seated	pulse_automated	pulse_manual	sysbp	trunkfat_%	waistc
armfat_percent	1(0)	-0.14(0.02)	-0.10(0.03)	0.14(0.03)	0.04(0.07)	0.12(0.08)	0.949(0.004)	0.89(0.01)
height	-0.14(0.02)	1(0)	0.92(0.01)	-0.12(0.03)	-0.12(0.07)	-0.07(0.07)	0.06(0.03)	0.15(0.02)
height_seated	-0.10(0.03)	0.92(0.01)	1(0.0)	-0.06(0.03)	-0.02(0.07)	-0.06(0.08)	0.07(0.03)	0.16(0.03)
pulse_automated	0.14(0.03)	-0.12(0.03)	-0.06(0.03)	1(0.0)	0.99(0.14)	-0.05(0.08)	0.14(0.03)	0.11(0.03)
pulse_manual	0.04(0.07)	-0.12(0.07)	-0.02(0.07)	0.99(0.14)	1(0)	-0.36(0.24)	0.02(0.07)	-0.00(0.07)
sysbp	0.12(0.08)	-0.07(0.07)	-0.06(0.08)	-0.05(0.08)	-0.36(0.24)	1(0)	0.09(0.08)	0.13(0.09)
trunkfat_percent	0.949(0.004)	0.06(0.03)	0.07(0.03)	0.14(0.03)	0.02(0.07)	0.09(0.08)	1(0)	0.86(0.01)
waistc	0.89(0.01)	0.15(0.02)	0.16(0.03)	0.11(0.03)	-0.00(0.07)	0.13(0.09)	0.86(0.01)	1(0)

Table 3.2: Genetic correlation for UKB anthropometric traits, estimated using LDSC software package on the same individuals and filtered SNPs used for the 200k cohort GWAS.

3.4 Discussion

In this paper, we tested the utility of multi-trait methods for improving risk score prediction on a variety of traits from the UK Biobank. We found that PGS computed using GWAS estimates outperformed multi-trait methods in several of our experiments, and performed comparably when not.

The Gaussian mixture models we tested, which allowed various combinations of sparsity and independence between traits, tended to perform similarly to each other, but their relative performance to other methods varied dramatically. For example, they achieved higher PGS accuracy than MTAG when computing PGS for an automated pulse measurement using the weakly correlated height as an auxiliary trait, but lower accuracy than MTAG and the stratified model for predicting manual pulse from a small sample using automated pulse measurement (which had a much larger study size) as an auxiliary trait.

The stratified model tended to perform similarly or slightly worse than MTAG. Both tended to perform strongly when the traits had high genetic correlation, and the trait of interest was underpowered. The stratified model may have performed better had we included SNPs with MAF ¡ 1%, since rarer SNPs may have a genetic architecture that differs more drastically from the genome-wide average than relatively common SNPs. For this paper, we applied the same SNP QC filters as in Turley *et al.* (2018) to choose SNPs to use in the PGS, though their filter excludes rare SNPs with the rationale that such SNPs are likely to have a different distribution of effects than the rest of the genome, a key assumption of the MTAG model but not the stratified model. The stratified model may also suffer from increased estimation error for the genetic covariance and non-genetic covariance matrices,

due to the decreased number of SNPs per bin used to estimate these model parameters.

This is far from a comprehensive evaluation of multi-trait methods for computing PGS, particularly since each experiment was conducted only once per set of traits. One could measure PGS on many bootstrapped samples of individuals, in order to obtain confidence intervals on the PGS R^2 obtained by different methods. Another area for future work would be to quantify the change in PGS accuracy of multi-trait methods as a function of study size and genetic correlation. Future studies could also evaluate similar models for modeling effects on one trait in different populations, potentially using GWAS estimates from well-represented groups to boost accuracy when computing PGS for a population that is underrepresented in the GWAS literature.

Chapter 4

Neural network guided detection of thyroid eye disease from external photos

4.1 Introduction

Thyroid eye disease (TED) is a progressive autoimmune disease (Bahn, 2010) with prevalence estimated to be 0.5 percent of the general population (Wiersinga and Bartalena, 2002). Advanced cases are characterized by facial disfigurement, chronic orbital pain, intractable diplopia, and vision loss (Bahn, 2010; Sabini *et al.*, 2017). As textbooks typically depict only the severe manifestations, uninitiated clinicians often overlook TED's subtle features, and misdiagnosis of TED patients early in their disease course is common. These patients may wait months to years for an accurate diagnosis (Mellington *et al.*, 2017; Estcourt *et al.*, 2009), and lose valuable time (Menconi *et al.*, 2014) when interventions such as behavior

modification and pharmacologic therapy (Bartalena *et al.*, 2017; Smith *et al.*, 2017; Winn and Kersten, 2021) might still be able to alter the disease course. Tools that accurately identify TED and direct patients to specialist care are would be of great value in expediting care for early and milder cases of TED.

Deep learning methods have been used to develop similar tools in other areas(Shen *et al.*, 2017), such as malignant breast cancer histopathology (Spanhol *et al.*, 2016; Hameed *et al.*, 2020), tuberculosis identification in chest radiography (Lakhani and Sundaram, 2017), and melanoma (Haenssle *et al.*, 2018; Brinker *et al.*, 2019) and other skin lesions (Esteva *et al.*, 2017; Liu *et al.*, 2020) in external photographs. These methods have also been applied to ophthalmic imaging, in the detection of glaucoma (Li *et al.*, 2018; Phene *et al.*, 2019), diabetic retinopathy (Gulshan *et al.*, 2016; Ting *et al.*, 2017; Raman *et al.*, 2019; Oh *et al.*, 2021), and age-related macular degeneration (Lee *et al.*, 2017; Grassmann *et al.*, 2018; Peng *et al.*, 2019) from fundus and optical coherence tomography (OCT) images. Given that patients with TED display external features characteristic of the disease, deep learning methods are particularly well-positioned to detect this orbital condition.

In this study, we develop a deep-learning based classifier to detect TED from external photographs and evaluate its effectiveness. This technology could be applied in the future to identify patients that would benefit from early treatment and follow up.

4.2 Methods

4.2.1 Data acquisition and labeling from clinical records

Photographic and clinical data were collected from patients evaluated by an experienced orbital specialist at the Stein Eye Institute (SEI) UCLA, a tertiary care, university-based orbital and ophthalmic plastic surgery practice. The associated Structured Query Language (SQL) database (McCann Medical Matrix, St. Louis, MO) was queried in order to identify two cohorts of patients: (1) patients with a confirmed clinical and radiologic diagnosis of TED and (2) a control cohort of patients with no evidence of TED. The second group was drawn from a cohort of patients presenting with epiphora or other lacrimal system problems, and those with facial rhytides and/or other cosmetic concerns. Records from January 1998 to October 2018 were screened.

The medical record of each patient identified from the query was reviewed to confirm the presence or absence of TED and assign ground truth labels. Patients with non-TED orbital conditions were excluded from both groups. Patients were also excluded if their identity or diagnosis could not be confirmed by review of the medical record.

Photographs from the date of each patient's initial consultation were collected and screened. Photos were excluded if they did not depict at least both eyes, eyelids, canthi and brows, the forehead, both temples, the glabella and the nasal dorsum. Photos were additionally excluded if poorly focused, if both eyes were not open, or any medical devices were visible in the frame. For patients in the TED cohort, photographs captured after orbital decompression or other TED-related oculofacial reconstructive surgery were excluded.

An additional test set was prepared from photographs and clinical data of TED patients

evaluated in a distinct clinic, the Doheny Eye Institute (DEI) UCLA. An alternative image repository (Axis Image Management, Sonomed Escalon, Lake Success, NY) and electronic medical record (Epic Systems, Madison, WI) were used to collect and label this dataset. In this set, images from both the first and second patient visits were included. Photos were screened using the inclusion and exclusion criteria as above. Additional data regarding TED grade and stage at the time of the initial visit was collected for these patients.

4.2.2 Image preprocessing

For each patient in the SEI dataset, a single front facing photograph with the patient's gaze in primary position was used. Where the patient demonstrated strabismus, photographs were selected in which at least one eye was in primary position. For each patient in the DEI dataset, a single front facing photograph was selected in a similar fashion for the first and second visit.

Each image was preprocessed by cropping to a region centered around the eyes, from above the eyebrows to below the lower eyelids, including part of the nasal dorsum and both temples. Eye detection was based on automated detection of face landmarks using the 5-landmark model from dlib (King, 2009) or a Haar cascade (Viola and Jones, 2001). The images were then scaled to 280 x 460 pixels, padding the cropped image with black pixels to achieve the desired aspect ratio if necessary. This preprocessing was intended to minimize potential confounding factors and speed up training by excluding irrelevant areas of the image.

4.2.3 Deep learning ensemble model training and evaluation

The model in this paper is an ensemble model, creating by performing a cross-validation where five neural networks with the same parameters were trained on different subsets of the SEI data, then using the predictions from the five component models to generate a final prediction. The ensemble creation process, model training, and ensemble output are illustrated in Figure 4.1.

The SEI dataset images were randomly split into 85% training and validation, and 15% test images. The 85% training and validation set was then used to form datasets for the cross-validation, where folds were created by further splitting this into 70% training and 15% as validation data with a different non-overlapping validation set for each fold. A neural network was trained separately on each fold by using the training dataset to update the weights of the network, and the validation dataset to monitor the training progress of the model to prevent overfitting (Caruana *et al.*, 2000). The test dataset was held out and used to evaluate final model performance. The DEI images were also held-out for training, in order to assess model generalizability to data from a different clinic and to assess model performance on patients classified by disease stage and grade. This is illustrated in Figure 4.1a.

Each deep learning model in the ensemble was a residual neural network (He *et al.*, 2015) with 18 layers (ResNet18), implemented in the python programming language using PyTorch. The neural networks were pretrained on ImageNet, a dataset of 1.4 million general images for the task of object recognition, then finetuned on the SEI training and validation datasets for the task of classifying images as TED or non-TED. This pretraining and finetuning strategy is commonly used for image classification tasks, and allows neural networks to be trained for

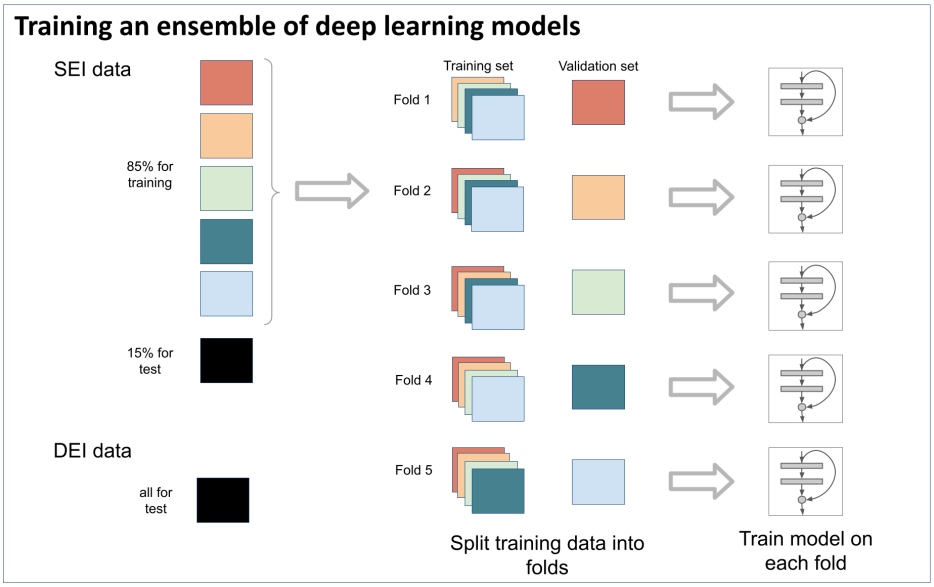
specific tasks where limited labeled data is available. The objective function was weighted cross-entropy loss with weights (1, 1.8) on the control and TED classes respectively, to reflect the imbalanced class sizes used during training. The optimizer was an Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 5×10^{-4}, paired with a scheduler to reduce learning rate on plateau, with a reduction factor of 0.25 and patience of 5 epochs. Each component network was trained with a batch size of 16 images. Training was halted if there was no improvement to the validation loss after 15 epochs, and the model reverted to its state with the best validation loss for testing.

Data augmentation was applied to the training data to ensure the models ignored irrelevant features in the image, such as lighting or precisely how much of the face is present in the image. This was done by selecting a random subset of transforms that include flipping the image horizontally, rotating by up to 10 degrees, distorting to simulate small changes in perspective, jittering brightness, jittering color, and taking a random 224 x 448 crop from the 280 x 460 preprocessed image. For each epoch and for each training image, a different set of these transforms was applied. Finally, the mean and standard deviation of each channel in the image were normalized in accordance with the normalization settings used for the initial training on Imagenet. When evaluating the model on the validation or test datasets, each preprocessed image was center cropped to 224 x 448 pixels and normalized as above, but other transforms were not applied. This is illustrated in Figure 4.1b.

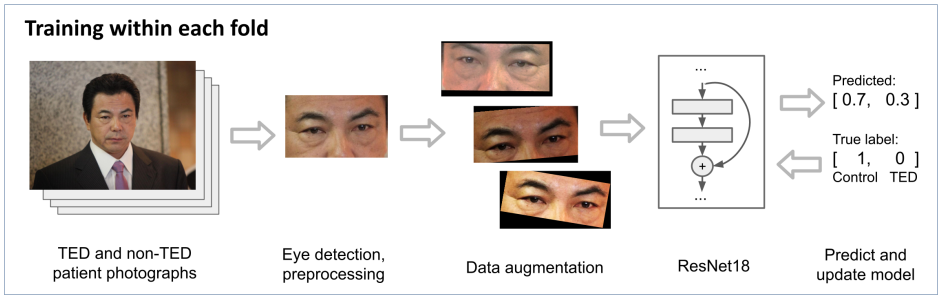
Each component neural network outputs prediction probabilities for the non-TED and TED classes, which correspond to that model's confidence in the prediction for each sample. To generate the final output for each test sample, the ensemble model then takes the output with the highest prediction probability for TED out of the five model outputs as its output.

To obtain a binary prediction of TED or non-TED for an image, the class probability for the TED class is thresholded. Hence the ensemble model will classify an image as TED if at least one of the component models classified the image as TED. This is illustrated in Figure 4.1c. For the accuracy metrics, a threshold of 0.5 was used.

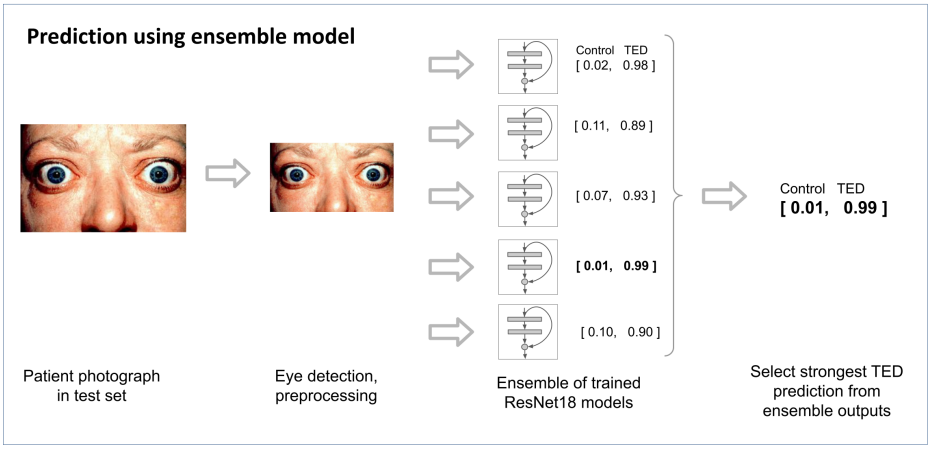
All experiments were performed on a PC equipped with a Nvidia Titan RTX graphics card with 32GB GPU memory.



(a)



(b)



(c)

Figure 4.1: Schematics illustrating the (a) cross-validation process used to create the ensemble of neural nets, (b) training process for each neural net, and (c) ensemble prediction. The validation set (different for each fold) is used to decide when to stop training. In the evaluation setting, the images are not augmented before being passed to the model. Example photographs in schematic from Route246 (2010) and Trobe (2011).

4.3 Results

4.3.1 Model performance on SEI dataset

An ensemble of neural networks model was used to classify TED from external photos after being trained on a dataset of 1,252 control images and 692 case images from SEI. It was evaluated on a held-out test set of 344 images from SEI. The model achieved an overall accuracy of 89.2% on the test dataset and recall (proportion of true cases which were correctly predicted, equivalent to true positive rate or sensitivity) of 93.4% (Table 4.1). Confusion matrices showing breakdown by class are shown in Figure 4.2. The receiver-operating characteristic (ROC) curve and precision-recall curve are shown in Figure 4.3. It is possible to use our model to achieve higher recall at the expense of lower precision (proportion of predicted cases which are true cases) by lowering the threshold at which a patient is classified as TED. For example, when the test precision is 50% (i.e. if the classification threshold were set such that only half of predicted TED patients actually had TED), the model achieves 99.2% recall on the test data, as shown in Figure 4.3b.

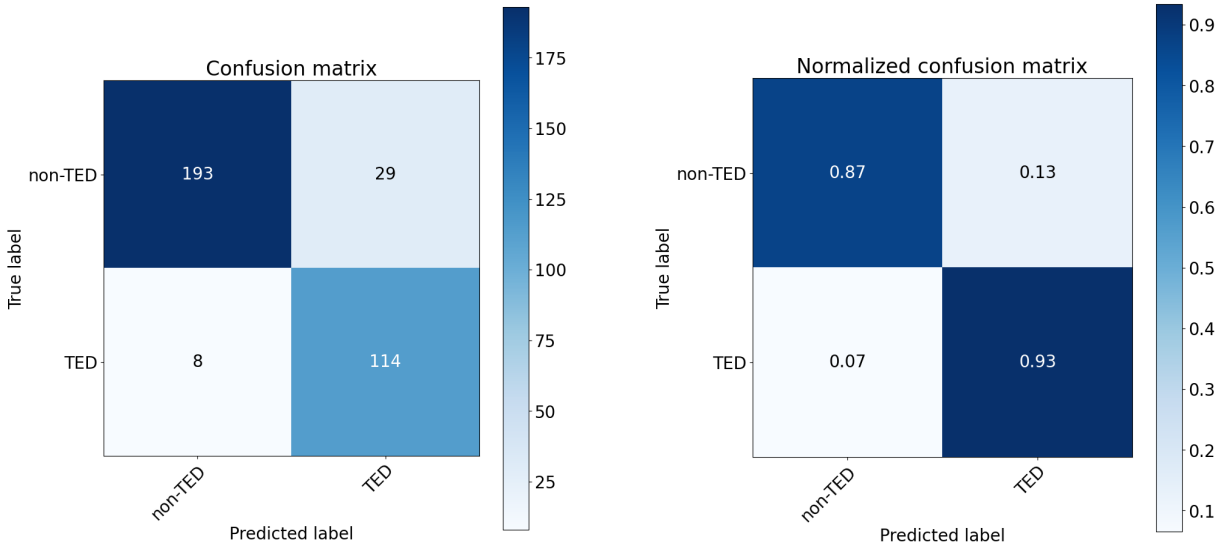
By design, the ensemble model achieves higher recall than any individual model in the ensemble, since it will predict an image as TED if any model in the ensemble predicts it as TED. The performance of the neural networks within the ensemble are also given in Table 4.2. The performance was similar across models on the SEI data, with a mean accuracy of $90.4 \pm 0.01\%$ and recall of $86.1 \pm 0.08\%$, though the precision ranged from 81.1% to 88.5%.

Table 4.1: TED classification model metrics. Model was trained on patient photos taken as part of clinical care at Stein Eye Institute (SEI). Model was evaluated on a held-out test set from SEI after training was complete. The model was also evaluated on two additional sets of images from a separate clinical location at Doheny Eye Institute (DEI), taken at either a patient's first or second visit to DEI. The DEI dataset contained only TED patients.

Dataset	n_{total}	n_{TED}	Accuracy	Specificity	Recall	Precision	F1 score
SEI test	344	122	0.892	0.869	0.934	0.797	0.860
DEI first visit	123	123	-	-	0.919	-	-
DEI second visit	99	99	-	-	0.919	-	-

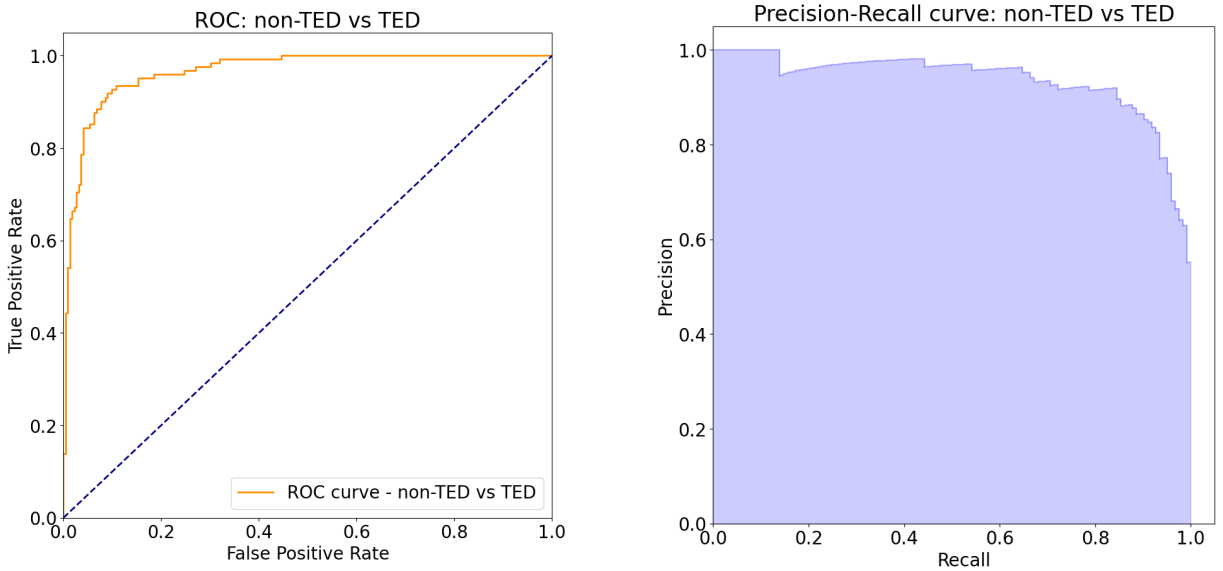
Table 4.2: Performance of individual models within ensemble. Table shows performance of the component models in the ensemble on the SEI test set, trained on different folds from a cross-validation.

Fold	Accuracy	Specificity	Recall	Precision	F1 Score
1	0.898	0.905	0.861	0.885	0.837
2	0.898	0.946	0.850	0.811	0.892
3	0.907	0.941	0.866	0.844	0.888
4	0.901	0.937	0.857	0.836	0.879
5	0.916	0.973	0.872	0.811	0.943
Mean	0.904	0.941	0.861	0.838	0.888
St. dev.	0.007	0.024	0.008	0.030	0.038



(a) (b)

Figure 4.2: Confusion matrices for TED classification model on SEI test data. (a) Counts for each class and predicted class. (b) Counts normalized by class size.



(a) (b)

Figure 4.3: ROC and precision-recall curve for TED classification model on SEI test data. (a) Receiver-operating characteristic curve (ROC), plotting recall (true positive rate) against specificity (true negative rate). (b) Precision-recall curve, plotting precision (proportion of predicted cases which are true cases) against recall.

4.3.2 Model performance on DEI dataset, stratified by TED stage and severity

To assess whether the model would generalize to images from sites not used in the training dataset, the model was also evaluated on photos of TED patients from DEI taken at the patient's first or second visit. The DEI dataset included patient information regarding disease stage for the first visit, noted by the attending orbital surgeon at time of photograph capture. Non-TED images were not collected for this dataset, so only recall is reported.

The model achieves recall rates of 91.9% on patient photos from the first and second DEI visit respectively, compared to 93.4% on the SEI dataset (Table 4.1). Stratifying the first visit data by stage and grade reveals that model recall increases with disease severity. The model achieves a recall of 86.8% on cases labeled as mild ($n = 68$), compared to 98.2% on cases graded from mild-moderate to severe ($n = 55$) (Table 4.3). The model recall is also higher for cases in the active or late active inflammatory stage, with recall of 98.3% ($n = 60$) compared to cases in the stable stage, with a recall of 85.7% ($n = 63$). These differences become more stark when looking at the performance of the individual models within the ensemble, which have a mean recall of $69.7 \pm 6.3\%$ for mild cases versus $92.0 \pm 3.0\%$ for mild-moderate and above; and $86.7 \pm 4.9\%$ for active and late-active versus $73.0 \pm 4.0\%$ in the stable stage. The variance in recall across component models on the DEI dataset is higher than the SEI dataset (standard deviation of 4.2% on 123 cases vs 0.8% on 122 cases) and particularly high for the mild and mild-moderate cases. Yet despite decreased consistency in individual model performance on the DEI dataset, the ensemble model achieved similar recall on the SEI and DEI datasets. This highlights the robustness of this ensemble approach

Table 4.3: Ensemble model performance stratified by TED inflammatory stage and grade. After being trained on images from SEI, the model was evaluated on images of TED patients at DEI from their first visit, where the stage and grade of the disease were noted by attending physician. The mean and standard deviation of recall for the component models in the ensemble are also given.

Category	$n_{predicted}$	n_{TED}	Recall	Recall per CV fold
By grade				
Mild	59	68	0.868	0.697 ± 0.063
Mild-moderate	11	11	1.000	0.927 ± 0.076
Moderate	32	33	0.970	0.891 ± 0.035
Moderate-severe	6	6	1.000	1.000 ± 0.000
Severe	2	2	1.000	1.000 ± 0.000
Severe (optic neuropathy)	3	3	1.000	1.000 ± 0.000
By stage				
Active	43	44	0.977	0.827 ± 0.061
Late active	16	16	1.000	0.975 ± 0.034
Stable	54	63	0.857	0.730 ± 0.040
Total	113	123	0.919	0.797 ± 0.042

compared to any single model within the ensemble.

4.4 Discussion

The deep learning-based ensemble model presented in this investigation achieved an accuracy of 89.2% and recall of 93.4% in the binary classification of TED and non-TED external images when evaluated on a 344 patient heldout test set. A total of 1,252 control and 692 case images taken from retrospective data from a clinical practice were used to train the ensemble of neural networks used by the model. We also evaluated the model on a set of 123 TED patient images taken from a clinic site not seen during trained, and achieved a recall rate of 91.9%, with active stage and severe grade associated with higher recall.

Deep learning has been used previously to screen for TED using computed tomography (CT) images (Wu and Zou, 2018). Such models hold potential to enhance diagnostic accuracy for patients suspected of having TED. However radiologic imaging has limitations as a screening tool, mostly related to cost, availability and the obligate exposure to ionizing radiation. With digital photography, the necessary equipment is readily available in first contact clinician offices, and the cost of image capture and storage is negligible.

There are limitations of the deep learning TED classifier developed in this study, many of them due to constraints on the available data. The current framework classifies patients based on a single front facing, primary position photograph. Photographs captured at different angles could provide additional information for model training. Additionally, clinical features in conjunction with image data have been demonstrated to increase deep learning classifier performance relative to image data alone in other disease contexts (Haenssle *et al.*, 2018). For TED, motility tests, clinical features, laboratory tests, and patient descriptions of non-external symptoms such as pain are all informative of disease

status. Future studies could incorporate this additional information as features in the model. A prospective approach to data collection would address this by collecting images at multiple angles and motility test results for all patients visiting a clinic, regardless of case-control status or disease severity, as well as standardizing the format in which motility results and symptoms are reported in the patient record for ease of extraction.

Great care is needed when training and evaluating deep learning models for high-stakes domains such as medical image diagnostics. The presence of confounding features for classification alongside relevant disease features can bias deep learning models. For example, in one study, models trained to detect pneumonia from chest radiographs learned to use ancillary marks on the film itself, rather than the disease's radiographic features, to make its decision (Zech *et al.*, 2018). Even when the marks were excluded from training, these models tended to perform worse on radiographs collected from other hospitals, versus on radiographs from the same hospital used for training, suggesting less obvious confounders or imperceptible image quality differences between hospitals (Zech *et al.*, 2018). We evaluated recall on images from a clinical site not seen during training (DEI) to assess generalizability. In this analysis, slightly lower recall was noted (93.4% vs 91.1%), with a larger performance gap when considering individual models within the ensemble. This gap could have been due in part to differences in patient composition or severity between the two sites, since the model achieved comparable or higher recall on more severe cases from the second site, and on patients showing active disease, particularly late-active stage disease.

One limitation of the model when applied to the DEI dataset, is that it did not perform as well on patients with mild disease. As interest in computational medicine grows in the orbital and oculoplastic surgery communities, future work could involve collaborations and

data pooling from multiple institutions and departments. This would improve our ability to train and evaluate models on large and diverse sets of data. Combined with the addition of clinical features and other patient data as features, we anticipate future iterations of TED classifiers can improve for all stages of disease.

In its present iteration, the model is not able to detect diseases other than TED, as this model was trained on a background of control patients without orbital disease. In future iterations of this deep learning based classifier, adding photographs of patients with other systemic, orbital, eyelid and ophthalmic conditions would facilitate development of methods to assess for many orbit and eye conditions simultaneously.

A key advantage of using deep learning as a screening or diagnostic tool is that it does not rely on expensive, invasive testing such as radiological or laboratory evaluation. This, in turn, is a key prerequisite in reducing barriers to accessing healthcare and shortening the time to specialist evaluation, accurate diagnosis and targeted therapy.

Chapter 5

Conclusion

My thesis work comprises three projects on computational methods for biology and medicine, specifically on modeling GWAS data from multiple traits, and on image recognition for disease. In Chapter 2, I present work on combining heterogeneous studies to increase power to detect relevant genetic variants from association studies, and in Chapter 3, comparing models for genetic effect sizes across studies in terms of their predictive power. In Chapter 4, I present a deep learning-based classifier for identifying patients with thyroid eye disease without the need for invasive testing or specialized imaging equipment, which could allow primary care physicians or even the patients themselves to recognize the disease more quickly and seek treatment.

As said by an emperor Marcus Aurelius around 100 or 200 AD, "time is like violent stream; as soon as a thing has been seen, it is carried away and another comes in its place, and this too will be carried away." The field of biology has been through many changes since then. Over the last twenty years since their inception, GWAS have not always lived up to their early hype. However, leaps and bounds were made in terms of improving basic GWAS

with strategies such as pooling multiple studies (one of the focuses of this thesis), larger sample sizes, and better correcting for confounders. Many new sequencing and data collection technologies have also been developed, including ATAC-seq, bisulfite sequencing, CRISPR-based screens, and single-cell versions of such technologies. Our resources for understanding human biology continue to grow rapidly beyond genotype data alone.

In the future, we may also see a variety of pretrained image recognition models specific to medicine, and the availability of public datasets to supplement training of machine learning models for medicine continues to grow. Some of public datasets already available include ophthalmologic fundus and iris images (Khan *et al.*, 2021) and dermatological images, e.g. (Tschandl *et al.*, 2018; Rotemberg *et al.*, 2021). In parallel, a wide range of deep learning models have been developed for image recognition which can be applied to medical imaging and other clinical data. Deep learning models developed for non-biological tasks can be translated to use on biological data, with NLP models being of particular relevance to parse biological sequence data or medical records. The creation of new technology and increasing data availability will constantly provide new challenges in biology to address using statistical and computational methods.

Bibliography

Andreassen, O. A., Harbo, H. F., Wang, Y., *et al.* (2014). Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. *Molecular Psychiatry*, **20**(2), 207–214.

Bahn, R. S. (2010). Graves' ophthalmopathy. *New England Journal of Medicine*, **362**(8), 726–738.

Bartalena, L., Veronesi, G., Krassas, G. E., *et al.* (2017). Does early response to intravenous glucocorticoids predict the final outcome in patients with moderate-to-severe and active graves' orbitopathy? *Journal of Endocrinological Investigation*, **40**(5), 547–553.

Berndt, S. I., Camp, N. J., Skibola, C. F., *et al.* (2016). Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nature Communications*, **7**, 10933.

Brinker, T. J., Hekler, A., Enk, A. H., *et al.* (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, **113**, 47–54.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., *et al.* (2015a). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, **47**(11), 1236–1241.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., *et al.* (2015b). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, **47**(3), 291–295.

Bycroft, C., Freeman, C., Petkova, D., *et al.* (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203–209.

Cai, N., Bigdeli, T. B., Kretzschmar, W., *et al.* (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature*, **523**(7562), 588–591.

Caruana, R., Lawrence, S., and Giles, L. (2000). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS 00, page 381–387, Cambridge, MA, USA. MIT Press.

Chen, L., Ge, B., Casale, F. P., *et al.* (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, **167**(5), 1398–1414.e24.

Chesler, E. J., Lu, L., Shou, S., *et al.* (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, **37**(3), 233–242.

Consortium, T. I. H. (2005). A haplotype map of the human genome. *Nature*, **437**(7063), 1299–1320.

Cross-Disorder Group of the Psychiatric Genomics Consortium, T. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, **381**(9875), 1371–1379.

Del-Aguila, J. L., Fernández, M. V., Schindler, S., *et al.* (2018). Assessment of the genetic

architecture of alzheimer's disease risk in rate of memory decline. *Journal of Alzheimer's Disease*, **62**(2), 745–756.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55**(4), 997–1004.

Dorn, G. W. and Cresci, S. (2009). Genome-wide association studies of coronary artery disease and heart failure: where are we going? *Pharmacogenomics*, **10**(2), 213–223.

Duong, D., Zou, J., Hormozdiari, F., *et al.* (2016). Using genomic annotations increases statistical power to detect eGenes. *Bioinformatics*, **32**(12), i156–i163.

Duong, D., Gai, L., Snir, S., *et al.* (2017). Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eqtls and increase the number of egenes. *Accepted*.

Duong, D., Gai, L., Uppunda, A., *et al.* (2020a). Annotating gene ontology terms for protein sequences with the transformer model. *biorxiv preprint*.

Duong, D., Uppunda, A., Gai, L., *et al.* (2020b). Evaluating representations for gene ontology terms. *biorxiv preprint*.

Eskin, E. (2015). Discovering genes involved in disease and the mystery of missing heritability. *Communications of the ACM*, **58**(10), 80–87.

Estcourt, S., Hickey, J., Perros, P., *et al.* (2009). The patient experience of services for thyroid eye disease in the united kingdom: results of a nationwide survey. *European Journal of Endocrinology*, **161**(3), 483–487.

Esteva, A., Kuprel, B., Novoa, R. A., *et al.* (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, **542**(7639), 115–118.

Evans, L. M., Tahmasbi, R., Vrieze, S. I., *et al.* (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, **50**(5), 737–745.

Fleiss, J. (1993). Review papers : The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, **2**(2), 121–145.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics*, **9**(5), e1003486.

Furlotte, N. A. and Eskin, E. (2015). Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics*, **200**(1), 59–68.

Gai, L. and Eskin, E. (2018). Finding associated variants in genome-wide association studies on multiple traits. *Bioinformatics*, **34**(13), i467–i474.

Grassmann, F., Mengelkamp, J., Brandl, C., *et al.* (2018). A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*, **125**(9), 1410–1420.

Grove, J., Ripke, S., Als, T. D., *et al.* (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics*, **51**(3), 431–444.

Gulshan, V., Peng, L., Coram, M., *et al.* (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, **316**(22), 2402.

Haenssle, H., Fink, C., Schneiderbauer, R., *et al.* (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, **29**(8), 1836–1842.

Hameed, Z., Zahia, S., Garcia-Zapirain, B., *et al.* (2020). Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors*, **20**(16), 4373.

Han, B. and Eskin, E. (2012). Interpreting meta-analyses of genome-wide association studies. *PLoS Genetics*, **8**(3), e1002555.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, **abs/1512.03385**.

Hu, Y., Lu, Q., Liu, W., *et al.* (2017). Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLOS Genetics*, **13**(6), e1006836.

Hyde, C. L., Nagle, M. W., Tian, C., *et al.* (2016). Identification of 15 genetic loci associated with risk of major depression in individuals of european descent. *Nature Genetics*, **48**(9), 1031–1036.

Jones, H. J., Heron, J., Hammerton, G., *et al.* (2018). Investigating the genetic architecture of general and specific psychopathology in adolescence. *Translational Psychiatry*, **8**(1).

Joo, J. W. J., Kang, E. Y., Org, E., *et al.* (2016). Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure. *Genetics*, **204**(4), 1379–1390.

Kamatani, Y., Matsuda, K., Okada, Y., *et al.* (2010). Genome-wide association study of hematological and biochemical traits in a japanese population. *Nature Genetics*, **43**(3), 210–5.

Kang, H. M., Sul, J. H., Service, S. K., *et al.* (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**(4), 348–354.

Khan, S. M., Liu, X., Nath, S., *et al.* (2021). A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, **3**(1), e51–e66.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, **10**, 1755–1758.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lakhani, P. and Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, **284**(2), 574–582.

Lam, M., Trampush, J. W., Yu, J., *et al.* (2017). Large-scale cognitive GWAS meta-analysis reveals tissue-specific neural expression and potential nootropic drug targets. *Cell Reports*, **21**(9), 2597–2613.

Lee, C. S., Baughman, D. M., and Lee, A. Y. (2017). Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmology Retina*, **1**(4), 322–327.

Lee, J.-Y., Lee, B.-S., Shin, D.-J., *et al.* (2013). A genome-wide association study of a coronary artery disease risk variant. *Journal of Human Genetics*, **58**(3), 120–126.

Li, Z., He, Y., Keel, S., *et al.* (2018). Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*, **125**(8), 1199–1206.

Liu, G., Hu, Y., Jin, S., *et al.* (2016). Cis-eQTLs regulate reducedLST1gene andNCR3gene expression and contribute to increased autoimmune disease risk: Table 1. *Proceedings of the National Academy of Sciences*, **113**(42), E6321–E6322.

Liu, Y., Jain, A., Eng, C., *et al.* (2020). A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, **26**(6), 900–908.

Logue, M. W., Panizzon, M. S., Elman, J. A., *et al.* (2018). Use of an alzheimer's disease polygenic risk score to identify mild cognitive impairment in adults in their 50s. *Molecular Psychiatry*.

Maier, R. M., Zhu, Z., Lee, S. H., *et al.* (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature Communications*, **9**(1).

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., *et al.* (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, **9**(5), 356–369.

Mellington, F. E., Dayan, C. M., Dickinson, A. J., *et al.* (2017). Management of thyroid eye disease in the united kingdom: A multi-centre thyroid eye disease audit. *Orbit*, **36**(3), 159–169.

Menconi, F., Profilo, M. A., Leo, M., *et al.* (2014). Spontaneous improvement of untreated mild graves' ophthalmopathy: Rundle's curve revisited. *Thyroid*, **24**(1), 60–66.

Nikpay, M., Goel, A., Won, H.-H., *et al.* (2015). A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, **47**(10), 1121–1130.

Oh, K., Kang, H. M., Leem, D., *et al.* (2021). Early detection of diabetic retinopathy based

on deep learning and ultra-wide-field fundus images. *Scientific Reports*, **11**(1).

Pazokitoroudi, A., Wu, Y., Burch, K. S., *et al.* (2020). Efficient variance components analysis across millions of genomes. *Nature Communications*, **11**(1).

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, **32**(4), 381–385.

Peng, Y., Dharssi, S., Chen, Q., *et al.* (2019). DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, **126**(4), 565–575.

Phene, S., Dunn, R. C., Hammel, N., *et al.* (2019). Deep learning and glaucoma specialists. *Ophthalmology*, **126**(12), 1627–1639.

Postmus, I., Warren, H. R., Trompet, S., *et al.* (2016). Meta-analysis of genome-wide association studies of HDL cholesterol response to statins. *Journal of Medical Genetics*, **53**(12), 835–845.

Purcell, S., Neale, B., Todd-Brown, K., *et al.* (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**(3), 559–575.

Purcell, S. M., Wray, N. R., Stone, J. L., *et al.* (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*.

Qi, G. and Chatterjee, N. (2017). Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. *biorxiv*.

Raman, R., Srinivasan, S., Virmani, S., *et al.* (2019). Fundus photograph-based deep learning

algorithms in detecting diabetic retinopathy. *Eye (Lond)*, **33**(1), 97–109.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., *et al.* (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, **8**(1).

Route246 (2010). Kokonoe.jpg. *Wikimedia Commons*. Published under Creative Commons Attribution 3.0 Unported license.

Sabatti, C., Service, S. K., Hartikainen, A.-L., *et al.* (2008). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, **41**(1), 35–46.

Sabini, E., Leo, M., Mazzi, B., *et al.* (2017). Does graves' orbitopathy ever disappear answers to an old question. *European Thyroid Journal*, **6**(5), 263–270.

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, **19**(1), 221–248.

Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *The American Journal of Human Genetics*, **101**(5), 737–751.

Shieh, Y., Hu, D., Ma, L., *et al.* (2016). Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Research and Treatment*, **159**(3), 513–525.

Smith, T. J., Kahaly, G. J., Ezra, D. G., *et al.* (2017). Teprotumumab for thyroid-associated ophthalmopathy. *New England Journal of Medicine*, **376**(18), 1748–1761.

Solovieff, N., Cotsapas, C., Lee, P. H., *et al.* (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, **14**(7), 483–495.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016). Breast cancer histopathological image classification using convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE*, **8**(7), e65245.

Sul, J. H., Han, B., Ye, C., *et al.* (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genetics*, **9**(6), e1003491.

Ting, D. S. W., Cheung, C. Y.-L., Lim, G., *et al.* (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, **318**(22), 2211.

Trobe, J. (2011). Proptosis and lid retraction from graves' disease.jpg. *Wikimedia Commons*. Published under Creative Commons Attribution 3.0 Unported license.

Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, **5**(1).

Turley, P., Walters, R. K., Maghzian, O., *et al.* (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, **50**(2), 229–237.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., *et al.* (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, **97**(4), 576–592.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer*

Vision and Pattern Recognition. CVPR 2001. IEEE Comput. Soc.

Wiersinga, W. M. and Bartalena, L. (2002). Epidemiology and prevention of graves' ophthalmopathy. *Thyroid*, **12**(10), 855–860.

Winn, B. J. and Kersten, R. C. (2021). Teprotumumab: interpreting the clinical trials in the context of thyroid eye disease pathogenesis and current therapies. *Ophthalmology*.

Wu, C. and Zou, Y. (2018). Application of transfer learning in the recognition of TAO. In *2018 13th International Conference on Computer Science & Education (ICCSE)*. IEEE.

Zech, J. R., Badgeley, M. A., Liu, M., *et al.* (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, **15**(11), e1002683.

Zeggini, E. and Ioannidis, J. P. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics*, **10**(2), 191–201.

Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, **11**(4), 407–409.