Genetic Prediction of Myopia

Neema Ghorbani Mojarrad

PhD Thesis

2019

Under the supervision of Professor Jeremy A. Guggenheim and Dr Cathy Williams

Acknowledgements

I would like to extend my sincere thanks and appreciation to several people for their help during my PhD and the development of this thesis. I am of course, extremely grateful and indebted to my supervisors Jez Guggenheim and Cathy Williams, for their patience with me and the project, their support and direction, contribution to the research, and introduction to the academic world. Your guidance has expanded my interest in academia and taught me to try and push myself to be a rigorous, conscientious, and open-minded researcher.

I would like to thank my advisor Keith Meek and PGR secretary Sue Hobbs for their support over the duration of this PhD. I'd also like to acknowledge the other PhD students in office 1.18 and send my thanks to the current and former members of the myopia genetics research group: Yvonne Huang, Denis Plotnikov, Alfred Pozarickij, and Rupal Shah.

A special thanks to all my friends and colleagues for the feedback and support they have given me. Particular thanks go to Louise, Tom, Nikita, Melissa, Kiranjit, James, and Suzie - for the great experiences we've shared during my PhD, the time spent together at Cardiff University, and the conferences we've attended together. I would also like to thank my parents for their support and helping me survive through various stresses over the last few years. I'd also like to acknowledge my cat Loki, who I'm grateful lived long enough to get me through the stresses of my PhD by making me smile and laugh, rest in peace.

I'd also like to thank all participants and all the staff involved in the UK Biobank and ALSPAC cohorts for making this research possible. Last (but certainly not least), I'd like to thank the College of Optometrists for funding this project.

Summary

The number of people developing myopia and high myopia has increased in recent years, throughout Europe and Asia. This increased incidence is set to escalate further, with reports of nearly 50% of the population likely to develop myopia by 2050. Although investigations into what is causing this 'myopia boom' and its associated risk factors is currently ongoing, it is still not entirely clear why certain individuals are affected and not others. It is suspected that changes in the environment within recent years, such as a global push towards education, may be responsible. However there is a large body of evidence that demonstrates the complexity of the refractive error phenotype and human emmetropisation, with reports identifying more than 150 genetic loci associated with myopia, implying some people may have a genetic predisposition to the condition. The aim of this thesis was to investigate whether inspecting genetic predisposition to myopia would allow us to detect individuals at risk, and determine whether a genetic model to predict children at risk was feasible. This may then help identify individuals who would benefit more from early intervention, or more regular monitoring.

Initially, 149 genetic variants that reached genome-wide statistical significance in a GWAS for refractive error carried out by the CREAM consortium were used to create a 'genetic risk score' to assess the accuracy with which incident myopia could be predicted in children from the ALSPAC cohort. Analyses were also carried out for another predictor, namely the children's number of myopic parents. The results suggested that the number of myopic parents was a better predictor of refractive error and incident myopia than the genetic risk score (R^2 = 4.8% vs. 2.6%). This was likely due to several limitations in the genetic risk score. Notably, the results also demonstrated that these two predictors were largely independent, hence prediction accuracy improved when they were used together (R^2 = 7.0%).

To try and increase the accuracy of genetic prediction, I took advantage of the recently released genetic data from the UK Biobank cohort, for which a proportion (23%) of individuals also had ocular measurements taken. A genome wide association study (GWAS) was performed for autorefraction-measured refractive error in European individuals with both genetic and refractive data (N = 95,505), which replicated many loci previously shown to be associated with refractive error. A regression model to

impute refractive error in UK Biobank participants who did not undergo autorefraction measurement was also created (to improve the accuracy of the existing genetic risk score by means of running an additional GWAS analysis, thus expanding the effective sample size used in the creation of the genetic risk score). A multi-variable model was developed using age of onset of first spectacle wear, age and gender; the model fit was optimised objectively. The resultant model yielded an imputed refractive error that was moderately explanative (R^2 = 0.30) for the variance of 'true' (autorefraction-measured) refractive error, as judged in an independent sample. A GWAS for imputed refractive error was carried out in 287,448 European UK Biobank participants who were not amongst the 95,505 individuals included in the original GWAS for autorefraction-measured refractive error. The genetic correlation between the 2 traits (imputed refractive error vs. autorefraction-measured refractive error) was r_g = 0.92, which confirmed that the imputed refractive error phenotype was a good surrogate for the true phenotype.

Summary statistics from the 2 GWAS analyses described above were combined, along with GWAS summary statistics for educational attainment taken from a published study (www.SSGAC.org). Meta-analysis was performed using 'multi-trait analysis of genome-wide association summary statistics' (MTAG). The accuracy of the genetic risk scores in predicting refractive error was assessed in an independent sample of European adults (the ALSPAC mothers cohort). The best prediction accuracy was achieved by combining summary statistics for all 3 traits (autorefraction-measured refractive error, imputed refractive error, and educational attainment). The resultant genetic risk score explained 11.2% of the variance of refractive error, and demonstrated an area under the receiver operating characteristics curve (AUROC) of 0.67 and 0.75 for predicting any (≤-0.75D) and moderate (≤-3.00D) myopia, respectively. Participants from the ALSPAC mothers cohort in the top 10th percentile of the genetic risk score were found to be at 6-fold greater risk of developing high myopia (≤-5.00D) compared to the remainder of the sample. The accuracy of the genetic risk score was also tested in individuals of Asian, Chinese, and Black ancestry. Prediction accuracy was reduced by approximately 50% in Asian and Chinese individuals. Prediction accuracy was worse still in those of Black ethnicity.

List of Figures

List of Tables

Abbreviations

ALSPAC	Avon Longitudinal Study of Parents and Children
AOSW-inferred MSE	Age of Onset Spectacle Wear Inferred Mean Spherical Equivalent
AOSW norm MSE	Age of Onset Spectacle Wear Normalised Mean Spherical Equivalent
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operative Characteristic
Auto	Autorefraction
Autorefraction MSE	Autorefraction Measured Mean Spherical Equivalent
BiLEVE	Biobank Lung Exome Variant Evaluation
cM	Centimorgan
CNV	Copy Number Varaint
CREAM	Consortium for Refraction Error And Myopia
DNA	Deoxyribonucleic Acid
EduYears	Education Years; Years in Full Time Education
ELSPAC	European Longitudinal Study of Parents and Children
GCTA	Genome-Wide Complex Trait Analysis
GREML	Genome Based Restricted Estimated Likelihood
GRM	Generic Relatedness Matrix
GRS	Genetic Risk Score
GWAS	Genome Wide Association Study
h^2	Narrow-Sense Heritability
H^2	Broad-Sense Heritability
HRC	Haplotype Reference Consortium
HWE	Hardy-Weinberg Equilibrium
INDEL	Insertion of Deletion
LD	Linkage Disequilibrium
LMM	Linear Mixed Model
LOCO	Leave One Chromosome Out
MAE	Mean Average Error
MAF	Minor Allele Frequency
MSE	Mean Spherical Equivalent
MTAG	Multi-Trait Analysis of Genome Wide Association Studies
NMP	Number of Myopic Parents
Ortho-K	Ortho-Keratology
PCA	Principal Component Analysis
POAG	Primary Open Angle Glaucoma
QQ plot	Quantile-Quantile Plot
ROC	Receiver Operating Characteristic
SNP	Single Nucleotide Polymorphism
SSGAC	Social Sciences Genetic Association Consortium
WHO	World Health Organisation
YOB	Year Of Birth

Contents

1 General Introduction and Literature Review

1.1 Outline

This thesis aims to investigate the ability to genetically predict refractive error and myopia development, taking advantage of the recently released data for the UK Biobank cohort of over 500,000 individuals. This chapter will define the common terminology used within this thesis, and give an overview on the relevant literature, including the genetics of myopia and its prediction using genetic information.

Initially, descriptions of how refractive errors are classified will be given, followed by a brief overview of the prevalence of myopia (both current and projected), and a discussion of the most strongly implicated risk factors. After this, descriptions of genetic terminology and explanations of common principles in quantitative genetics that are relevant to this project will be outlined. Finally, a general overview of the myopia genetics literature will be presented, including the context of where the analyses in this thesis fit into the wider context of the management of patients with myopia.

1.2 Introduction to Refractive Error

1.2.1 Classification of Refractive Errors

1.2.1.1 Myopia

Myopia is defined as a form of refractive error (or ametropia) where the axial length of the eye is too long for the refractive power of its ocular components, causing light to focus in front of the retina i.e. the eye's corresponding focal length is shorter than the axial length (Millodot 2014). Thus, if this refractive error is left uncorrected, it results in blurry vision for the individual at far distances. Although this discrepancy between focal length and axial length can be caused by a disproportionately high power of the cornea or lens, most non-syndromic myopia is caused by excessive axial elongation of the eye (Grosvenor and Scott 1993; Morgan et al. 2012).

The refractive error threshold accepted to be the lower boundary for myopia categorisation has been widely debated (BHVI and WHO 2016). Some researchers have selected a threshold of ≤-0.50D (Rosenfield and Gilmartin 1998; Czepita et al. 2019;

Ueda et al. 2019), whereas others have classified myopia as a refraction of ≤-1.00D (Guggenheim et al. 2012; Cumberland et al. 2016). This is usually due to methodology, with studies involving self-reported information or younger participants with refractive error measured using non-cycloplegic autorefraction setting more stringent myopic thresholds in order to reduce misclassification. However, a recent meta-analysis indicated the most commonly used threshold is ≤-0.50D, and recommended that this threshold be applied for all future studies if there is no plausible risk of bias, such as those described above (Flitcroft et al. 2019).

1.2.1.2 High Myopia

High myopia is a subcategory of myopia, for which, again, there is no consensus regarding the threshold used for classification (BHVI and WHO 2016). A meta-analysis indicated that a threshold of ≤-6.00D or <-6.00D is most commonly used for defining high myopia, with 61% of reports using either of these thresholds (Flitcroft et al. 2019). Accordingly, ≤-6.00D has been proposed as the preferred refractive threshold by the International Myopia Institute (Flitcroft et al. 2019) with the recommendation that this level continue to be used for consistency, but that a threshold of ≤-5.00D may still be useful in certain circumstances. The ≤-5.00D threshold has also been recommended by other authors for its clinical relevance, as the unaided vision would typically be <3/60, matching the diagnostic threshold for blindness or severe sight impairment. The latter threshold was recommended by the World Health Organisation (BHVI and WHO 2016), suggesting the use of either threshold could be justified.

High myopia is sometimes due to monogenic (see Section 1.3.2) or syndromic conditions, for which a single faulty gene can cause early onset high myopia (Morgan et al. 2012). There are 261 syndromes listed in the Online Mendelian Inheritance in Man (OMIM) database for which myopia is a feature, however this form of myopia is usually much less common than other multi-factorial causes of myopia.

1.2.1.3 Hypermetropia/Hyperopia

Hypermetropia (or hyperopia) is the opposite state of myopia, in which the refractive power and corresponding focal length of the eye is longer than the axial length (Veerappan et al. 2009). Usually infants are born hyperopic, after which their ocular components begin to change to adapt to visual stimuli to correct any innate ametropia

(Flitcroft 2014). This process, termed emmetropisation, involves both a passive and active component (Mutti et al. 2005). The passive component involves an increase in eye size – specifically, a coordinated increase in axial length and flattening of corneal curvature – which is largely under genetic control (Wallman 1993). The active component involves a visually guided feedback mechanism in which the rate of axial elongation is fine-tuned to the clarity of the retinal image (Wallman 1993). Despite intense research, the mechanism by which active emmetropisation detects of the 'sign-of-defocus' (i.e. whether the visual image is focussed in front of the retina or behind the retina in an unaccommodating eye) is unclear. However, evidence that longitudinal chromatic aberration provides information on the sign of defocus has been obtained in a range of animal models (Rucker 2019).

As hyperopia does not carry the same associated comorbidities and secondary disease risks as myopia, it has not been as thoroughly investigated. However similarly to myopia, the threshold for defining hyperopia has varied from study to study, with no general consensus. Threshold levels of +0.50D (Yuan et al. 2015) and +1.00D (Cumberland et al. 2016) are common.

1.2.2 Secondary Conditions Associated with Myopia

Having myopia usually means that throughout their lifetime, an individual will require ocular correction to view objects at a distance. In countries with limited healthcare resources this can lead to a significant proportion of the population suffering from visual blur that would be classified as correctable visual impairment or blindness (Flaxman et al. 2017). Overall, uncorrected refractive error (with myopia being the greatest contributor) is the leading cause of moderate or severe visual impairment worldwide, and the second most frequent cause of blindness after cataract.

However, this is not the only or greatest concern regarding myopia. As myopia is primarily due to an increase in axial length, which is accompanied by stretching and thinning of the retina, choroid and sclera, individuals with myopia have an increased risk of many co-morbidities including primary open angle glaucoma (POAG), maculopathy, and retinal detachment (Marcus et al. 2011; Flitcroft 2012; BHVI and WHO 2016).

The risk of developing any of the above co-morbidities is greater in those with high myopia (Wong et al. 2014), however this is not to say that lower levels of myopia are

insignificant. Flitcroft (2012) reported that any level of myopia increases the risk of myopic maculopathy, posterior subcapsular cataract, POAG, and retinal detachment by 2-fold compared to someone without myopia. Another study argued that reducing myopia progression by 1.00D during childhood should reduce the incidence of myopic maculopathy by 40% (Bullimore and Brennan 2019).

In summary, uncorrected refractive error, particularly uncorrected myopia, is one of the most common causes of correctable visual impairment worldwide currently, and increases the risk of many serious secondary eye diseases that could cause loss of vision and leads to an increased burden for healthcare providers.

1.2.3 Prevalence of Refractive Error and Myopia

Estimates of the prevalence of myopia demonstrate high variability depending on the population studied, method of refractive measure performed (i.e. with or without the use of cycloplegia) and their ethnicity, and age. Most myopia prevalence studies have focussed on European and Asian populations. Moreover, these studies have used different thresholds for categorising myopia and high myopia, leading to difficulty in comparing between studies. In the section below describing current prevalence estimates, any study which used a cycloplegic has been specified, with the majority of prevalence studies not having used this.

1.2.3.1 Current Prevalence of Myopia

A meta-analysis of 62,000 European adults (98% with white European ancestry over the age of 25 years) estimated that 31% of the population were myopic, with just under 3% of the population having a refractive error ≤-6.00D (Williams et al. 2015b). A study in 2009 reported an overall prevalence of myopia within the white American population in 1999-2004 of 43% (Vitale et al. 2009). More locally, prevalence rates within the UK adult population also appear similar to that of the European sample with a UK twin-based cohort demonstrating 34% and 32% of the adults aged 50-54 and 55-59, respectively being myopic with a threshold of ≤-0.75D (Williams et al. 2013).

The prevalence of myopia (≤-0.50D; cycloplegic autorefraction) in UK children between the ages of 6 and 7 years old is between 2.8-5.7%, and increases to approximately 17.7-18.6% in 12-13 year-olds (O'Donoghue et al. 2010; Logan et al. 2011). As these children are still adolescents, it is likely that the number of individuals who will eventually

develop myopia will increase further. A 'cohort effect' has also been identified: the prevalence of myopia appears to be higher when investigating more recently-born individuals within large population samples of adults. For example, the E3 Consortium meta-analysis of European adults demonstrated that nearly 50% of those aged 25-29 years old had myopia of -0.75D or greater compared to 27% in those aged 55-59 years old (Williams et al. 2015b). However, other researchers have questioned the validity of such a cohort effect, instead arguing that myopia naturally reduces in prevalence with age (see below).

Prevalence estimates for myopia tend to be highest in young East-Asian populations, reaching epidemic levels (Morgan et al. 2018). Adult estimates in Singaporean, Chinese, Indian, and Malay adult males showed a myopia prevalence of 79%, 82%, 69%, and 65%, respectively (Wu et al. 2001), with over 10% of the Singaporean population demonstrating high myopia (<-6.00D). However, it may be argued that this was not a representative sample, as it comprised of conscripted military personnel, likely incorrectly weighted due to using 16-25 year olds (down-weighted as younger individuals may still go on to develop myopia later on, however potentially up-weighted compared to the population due to the young sample and lack of cycloplegia). A meta-analysis of Asian populations demonstrated approximately 28% of adults were myopic, using a threshold of ≤-0.50D (Pan et al. 2015).

The suggested cohort effect observed in Europeans has also been reported within Asian populations. A meta-analysis identified an increased prevalence of nearly 50% of adults aged below 29 years old being myopic, compared to 26% in those aged 30 to 39 years old (Pan et al. 2015). It should be noted that this meta-analysis included reports of more elderly populations (70+ years) that did not control for factors such as cataract development; consequently the authors reported a U-shaped relationship between myopia prevalence and age.

Furthermore recent prevalence estimates for Asians seem to be even higher in children; a study found that 88% of school children in China had myopia of ≤-0.50D when measured with non-cycloplegic autorefraction (Chen et al. 2018), with other reports from urban Chinese children also demonstrating a 65-80% prevalence with the same non-cycloplegic myopia threshold (He et al. 2004; You et al. 2014; Wu et al. 2015). Other

Asian countries have also shown high prevalence levels of myopia in young adult populations (using the same ≤ -0.50D threshold, but using cycloplegia) with 97% of 19 year olds in South Korea having myopia (Jung et al. 2012), and 84% and 21% of Taiwanese 16-18 year olds being myopic (≤ -0.25D) and highly myopic (≤ -6.00D), respectively, using cycloplegic autorefraction (Lin et al. 2004). Overall a consistent pattern in both European and Asian populations is that the prevalence of myopia is increasing as younger children and adolescents mature compared to previous generations, with a relatively higher prevalence in East Asian populations for the same age categories (Figure 1.1).

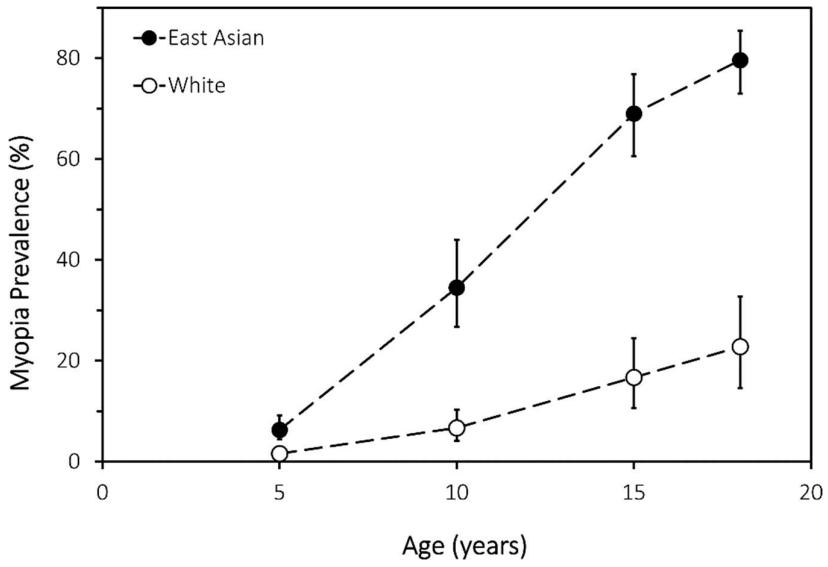


Figure 1.1 Prevalence of myopia by age for East Asian and White ethnicities from a meta-analysis in 2005. Error bars are 95% CIs. Taken from Wolffsohn et al. (2019) adapted from data by Rudnicka et al. (2016).

1.2.3.2 Future Prevalence of Myopia

As mentioned above, the evidence for an increasing prevalence of myopia in more recent birth cohorts has been questioned. A counter argument is that individuals naturally become less myopic as they get older, i.e. a longitudinal effect rather than a cohort effect (Mutti and Zadnik 2000). However, evidence continues to accumulate for a cohort effect. For example, in age-matched individuals from the United States, the myopia prevalence has increased over the last 30 years (Figure 1.2A) (Vitale et al. 2009). Moreover, there have been reports of similar findings for increased levels of myopia compared to previous estimates in the same populations (Williams et al. 2015b; Zhou et al. 2016; Chen et al. 2018; Morgan et al. 2018; Ueda et al. 2019). Williams et al. (2015a)

demonstrated a cohort effect when looking at individuals from a European population aged 40-79, with a higher myopia prevalence observed for more recent birth decades, as shown in Figure 1.2B.

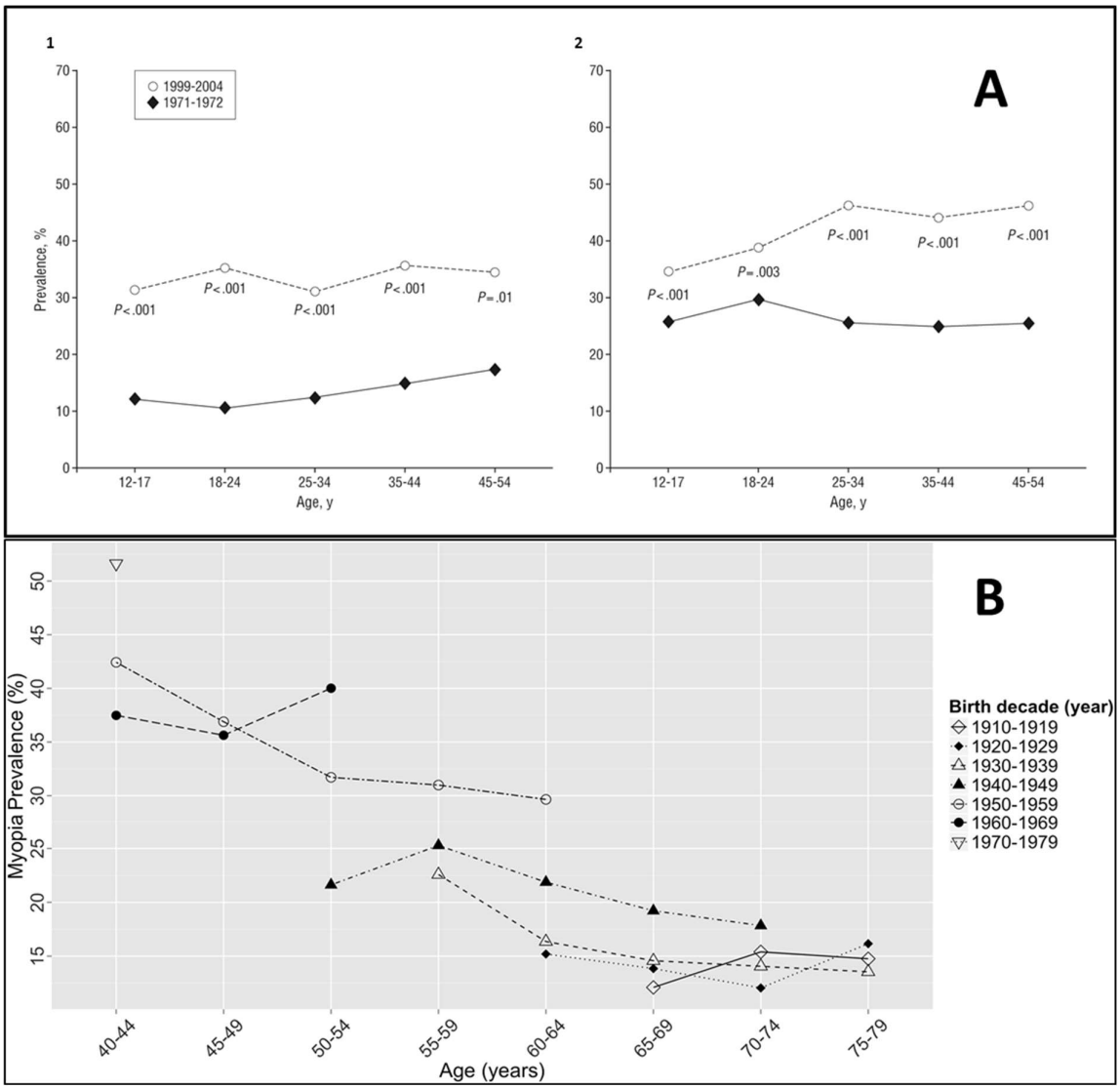


Figure 1.2 Graphical representations demonstrating the increase in myopia prevalence in recent birth cohorts. Panel A demonstrates the changes in prevalence between the early 1970's and the late 1990's in America, with subpanel (1) portraying the change in White individuals, and (2) portraying the change in Black individuals (taken from Vitale et al. (2009)). Panel B shows the prevalence of myopia in birth cohorts taken at different times and age ranges, demonstrating an increased prevalence in recent cohorts (taken from Williams et al. (2015a)).

As well as a general increase in myopia prevalence, there has been evidence that the age at which children become myopic has been decreasing i.e. on average children are becoming myopic at a younger age. In a study of Taiwanese children, Lin et al. (2004)

found that the average age children were becoming myopic had reduced from 11 to 8 years-old between 1983 and 2000. An earlier age of onset of myopia may lead in turn to an increase in the prevalence of high myopia, as children who develop myopia at younger ages tend to develop higher degrees of myopia, on average (Iribarren et al. 2009; Williams et al. 2013; Chua et al. 2016).

The increased prevalence of myopia in recent birth cohorts, as well as the reduced average age of onset, suggest that the prevalence of myopia and high myopia around the world are likely to increase in the future, with estimates of 50% and 10%, respectively, by 2050 (Holden et al. 2016). Given the association between myopia and secondary ocular disorders such as glaucoma and maculopathy (Flitcroft 2012), predicting children at an increased risk of developing myopia at an early age (prior to onset) would be beneficial and allow clinicians to monitor such at-risk individuals more closely and intervene to slow myopia progression at an early stage.

1.2.4 Environmental Influences and Risk Factors for Myopia

Given the rapid rise in myopia prevalence during the last 30 years (Lin et al. 2004; Vitale et al. 2009; Dolgin 2015), it has been argued that genetics cannot be directly responsible, because 30 years would only correspond to 1-2 generations, which is insufficient for temporal genetic change (Lim et al. 2014). Therefore changes in environmental risk factor exposure have been proposed as the primary cause of the recent increase in myopia prevalence (Holden et al. 2016). The environmental risk factors most widely studied are discussed below.

A relationship between educational attainment and myopia has been shown in a range of different study populations; a similar relationship between IQ and myopia has also been observed (Rosner and Belkin 1987; Au Eong et al. 1993; Saw et al. 2004; Morgan and Rose 2005; Pan et al. 2012; Williams et al. 2015a). For instance, adult Inuit populations had a myopia prevalence of 1.2% before the introduction of a formal education system (Lasker 1956), but within two generations, individuals under the age of 30 exhibited a prevalence of myopia of up to 58% (Young et al. 1969; Morgan et al. 1975). Evidence for a causal role of education in myopia has come from recent Mendelian randomisation studies, using genetic susceptibility as an instrumental variable (Cuellar-Partida et al. 2015; Mountjoy et al. 2018). This is the best current

evidence for a causal role of education, as randomised controlled trials for educational attainment would be unethical.

Reduced time spent outdoors is associated with an increased incidence of myopia (Guggenheim et al. 2012; French et al. 2013a; Guggenheim et al. 2014; He et al. 2015). In randomised controlled trials, increased time outdoors demonstrated a protective effect against myopia development (Ngo et al. 2014; He et al. 2015; Jin et al. 2015; Barry et al. 2016; Shah et al. 2017; Wu et al. 2018). The mechanism responsible has commonly been attributed to the increased light level outdoors: the so-called 'light-dopamine regulatory theory' (Witkovsky 2004; Rose et al. 2008; Smith et al. 2012; Hua et al. 2015). A crude explanation of this theory is that the release of dopamine in the retina - a neurotransmitter known to inhibit the rate of eye growth - is stimulated by bright light (McCarthy et al. 2007). This link of reduced light levels and increased rate of myopia can explain the effects of light deprivation studies, where animal models investigating myopia have shown increases in axial length and myopia with reduced light levels (Howlett and McFadden 2005; Ashby et al. 2009; Karouta and Ashby 2015). The importance of light levels being they key factor behind time outdoors' effect is further supported by the limited association of physical activity with myopia development (Rose et al. 2008; French et al. 2013a).

Studies investigating lighting changes and their potential interactions with circadian rhythms have also shown some significant findings. Reports using animal models have demonstrated a relationship of altered circadian rhythms and changes in lighting exposure during the day with an increased level of myopia development (Mutti et al. 1998; Norton and Siegwart 2013; Stone et al. 2013; Nickla and Totonelly 2016), leading to the suggestion that an altered, unnatural light cycle may have an impact on ocular development.

Furthermore, an investigation into disruptive night-time lighting in children found an association of ambient light at night and an increased rate of myopia (Quinn et al. 1999) suggesting a potential link for study, however this finding has not been replicated (Zadnik 2001; Guggenheim et al. 2003).

The relationship between insufficient time outdoors and myopia risk prompted investigation into vitamin D deficiency as a potential risk factor for myopia development.

Mutti and Marks (2011) measured serum vitamin D levels in a small sample of 13-25 year olds and concluded “myopes appear to have lower average blood levels of vitamin D than non-myopes”. Several subsequent cross-sectional studies have supported this association (Choi et al. 2014; Yazar et al. 2014; Tideman et al. 2016). However, Guggenheim et al. (2014) pointed out that vitamin D levels would be expected to be lower in children spending relatively less time outdoors. In a longitudinal study, these authors found no evidence that serum vitamin D mediated the relationship between time outdoors and myopia (Guggenheim et al. 2014). A review further deliberated on the lack of strong evidence for a causal link, as earlier studies which found associations failed to control for important confounders, such as time spent outdoors and sunlight exposure (Pan et al. 2017). Yet another study demonstrated a negative association of serum vitamin D levels and axial length in children, even after controlling for time outdoors; but it was not possible to infer whether this was due to a causal relationship or residual confounding (Tideman et al. 2016). However, a Mendelian randomisation study using genetic variants associated with vitamin D levels suggested that there was at most only a very small contribution from vitamin D on myopia development, indistinguishable from zero, and that the previous positive associations demonstrated would be likely due to confounding (Cuellar-Partida et al. 2017). Generally, the evidence for vitamin D being a risk factor for myopia appears to be limited, and could be considered a proxy measure for time outdoors, rather than an independent risk factor. Overall with regard to time outdoors, there is strong evidence from randomised controlled trials that time outdoors reduces the incidence of myopia, but the exact underlying mechanism remains unknown.

Near work (or increased time reading) has also been proposed as a risk factor for myopia development. A number of different factors relating to near work have been considered, including reading distance, posture, and length of time spent reading (Goss 2000; Hartwig et al. 2011; Lin et al. 2013). Because of this variation, it has been hard to compare across studies. A meta-analysis looking at reports for myopia and near work activities in children found an odds ratio of 1.14 for being myopic for every hour of increased near work per week (Huang et al. 2015). However, the results for this risk factor have been inconsistent (Mutti and Zadnik 2009), particularly regarding whether the effect of time outdoors has or has not been accounted for (Ip et al. 2008; Rose et al.

2008; Wu et al. 2013). This may suggest that increased near work could be a proxy for reduced time spent outdoors. It is currently unclear if the association between education and myopia (see above) is due to time engaged in near work during the school day.

Parental myopia has been shown consistently to be associated with an increased risk of myopia (Mutti et al. 2002; Saw et al. 2006; Jones-Jordan et al. 2010; O'Donoghue et al. 2015; Zadnik et al. 2015; Zhang et al. 2015b). However the precise degree of risk associated with having 0, 1, or 2 myopic parents varies from study to study. Mutti et al. (2002) reported an odds ratio for myopia of 3.31 and 7.29 in children with 1 or 2 myopic parents, respectively, compared to those with no myopic parents. However, Saw et al. (2006) reported odds ratios of only 1.63 and 1.70 for 1 and 2 myopic parents, respectively. Such discrepancies may be due to differences in the sample demographics, e.g. children of differing ethnicity. Age may also contribute to the incongruity, as French et al. (2013b) reported that having myopic parents was associated with an increased risk of myopia in 5-6 year olds, but not when the same children reached 13 years old.

The best current predictor of myopia risk is a 'pre-myopic' cycloplegic refraction of ≤+0.75D at the age of 6 years (Zadnik et al. 2015), which has an AUROC of 0.87. Zadnik et al. found that the addition of other risk measures, such as near work, time outdoors, or parental myopia offered minimal improvement in the accuracy of myopia prediction. However, the sensitivity and specificity of Zadnik et al.'s prediction model was lower at younger ages (aged around 6 years old) compared to older ages (aged around 11 years old) (Jones-Jordan et al. 2010; Zadnik et al. 2015), and was not evaluated in children aged under 6 years, meaning that the investigation of a predictor that could be used before the age of 6 may be useful. Moreover, another study using Chinese twins between the ages of 7 and 15 years old created multiple multi-variable models, which included ocular and genetic measures to assess prediction accuracy for high myopia (Chen et al. 2019). They found that age (both on its own and as a polynomial term), gender, parental spherical equivalent, and genetic risk score were significantly associated with myopia, but that the addition of the genetic risk score on top of age, gender, and measured refractive data did not enhance the predictive performance of their model (AUROC > 0.95). It should be noted that this model was used for predicting high myopia after the age of 13; by this age many children may already be myopic, and

therefore similarly to the study by Zadnik et al. the model may not be applicable to identifying at risk children who are already 'pre-myopic', i.e. on a refractive trajectory towards myopia.

1.2.5 Current Myopia Management and Myopia Control Strategies

There are currently two main approaches for slowing myopia progression that have been investigated: optical and pharmacological. The majority of studies investigating pharmacological interventions have tested the use of atropine eye drops. High (1%), medium (0.1-0.5%) and low (≤0.01%) dose atropine eye drops have all demonstrated a reduction in myopia progression compared to placebo or historical control groups (Chia et al. 2013; Chia et al. 2015; Huang et al. 2016). The 0.01% low dose atropine treatment demonstrated similar efficacy to that of high dose atropine, with less adverse effects such as reduction in accommodation and enlarged pupil size, along with a smaller 'rebound effect', whereby the reduction in myopia progression can accelerate post-treatment (Chia et al. 2012). However, it should be noted that low dose atropine has not shown the same consistent impact on reducing axial length growth alongside refractive error (Yam et al. 2019). Moreover, the use of atropine has been investigated largely in Asian populations (Chia et al. 2012; Kumaran et al. 2015; Huang et al. 2016), with investigations in white or European populations performed in smaller limited samples (Loughman and Flitcroft 2016; Polling et al. 2016).

Optical methods for reducing the progression of myopia that have demonstrated some success include multifocal/enhanced depth of focus lenses, defocus incorporated multiple segment (DIMS) spectacle lenses, and orthokeratology (ortho-K). Soft extended depth of focus lenses have been shown to reduce myopia progression by 10-60% over 2 years (Sankaridurg et al. 2019). A three year randomised control trial of a multifocal contact lens showed 59% and 52% relative reductions in the progression of myopia and axial elongation compared to control single vision contact lenses (Chamberlain et al. 2019). The DIMS spectacle lenses have also demonstrated similar efficacy, with a relative reduction of myopia progression and axial elongation by 52% and 62%, respectively (Lam et al. 2019). DIMS lenses incorporate a ring-shaped area containing tiny lenslets that provide additional plus power, around a 9mm diameter central zone that provides clear and central vision. Due to the change in corneal curvature and refractive error that

accompanies ortho-K lens wear, it is difficult to directly measure the reduction in myopia progression that occurs. However, ortho-K has been shown to reduce axial elongation, on average, by 43% relative to a control group (Cho and Cheung 2012). There have been few studies into possible rebound effects for ortho-K and multifocal lenses, although Cho and Cheung (Cho and Cheung 2017) suggested that axial elongation may resume upon cessation of lens wear (in teenagers). It is not possible to perform a double blind randomised controlled trial of ortho-K due to the fitting schedule.

Despite the progress made in interventions to reduce myopia progression, no optical or pharmacological intervention has been designed or tested with the aim of *preventing* incident myopia. As mentioned previously, increased time outdoors has been shown to prevent myopia onset in children, with randomised controlled trials of increased outdoor time in a school setting showing a reduction in the incidence of myopia (Wu et al. 2013; Barry et al. 2016). However, the evidence regarding time outdoors' efficacy in *slowing* myopia progression is not always consistent, with some studies showing no or limited effect of reducing myopia progression, and other studies demonstrating beneficial effects (Xiong et al. 2017; Cao et al. 2019).

In summary, although there have been advances in myopia intervention with regard to the reduction in progression (Huang et al. 2016; Chamberlain et al. 2019), and time outdoors with lower myopia incidence (Morgan et al. 2018), there is still more information required to fully understand and optimise current myopia interventions. This includes how to further improve efficacy, exploring why some people may not respond as well as others to the intervention, and to determine which intervention may be better suited to specific individuals based on their demographics or ocular status e.g. different level of myopia or different ethnicity.

1.3 Introduction to Quantitative Genetics and Complex Traits

1.3.1 Phenotypes, Genetic Variants, and Alleles

A phenotype is defined as a trait that is observable and measurable. This can be quantitative (for example height, which can be measured with a numerical value on a linear scale) or categorical (having two or more classes, for example eye colour). Phenotypes are determined by the action and possible interaction of genes and/or the

environment (see Section 1.3.4). The reason that humans can present with a range of different phenotypes within the same environment is because of variation within their genetic material i.e. having different genotypes. These differences can range from a single nucleotide substitution or an insertion or deletion of several nucleotides at a point in the genome (Frazer et al. 2009). Genetic variation has been estimated to occur in 0.6% of the human genome (1000 Genomes Project et al. 2015) with the most commonly occurring type of genetic variation being simple single nucleotide substitutions, termed single nucleotide polymorphisms (SNPs) (Wang et al. 1998). For example, at a specific locus in the genome, the majority of the population may have a 'G' nucleotide, however due to previous mutation some individuals may have inherited a 'T' nucleotide. These different nucleotides at polymorphic sites (places in the genome with frequently observed variation between individuals) are commonly referred to as alleles. Should a polymorphic site present with a minor allele frequency (MAF) of 1% or more i.e. the less common allele is present in more than 1% of the population, the polymorphism is classified as a 'common' (Wang et al. 1998). If the MAF is lower than 1%, the polymorphism is classified as a 'rare variant' or mutation. An example of a SNP allele is shown in Figure 1.3.

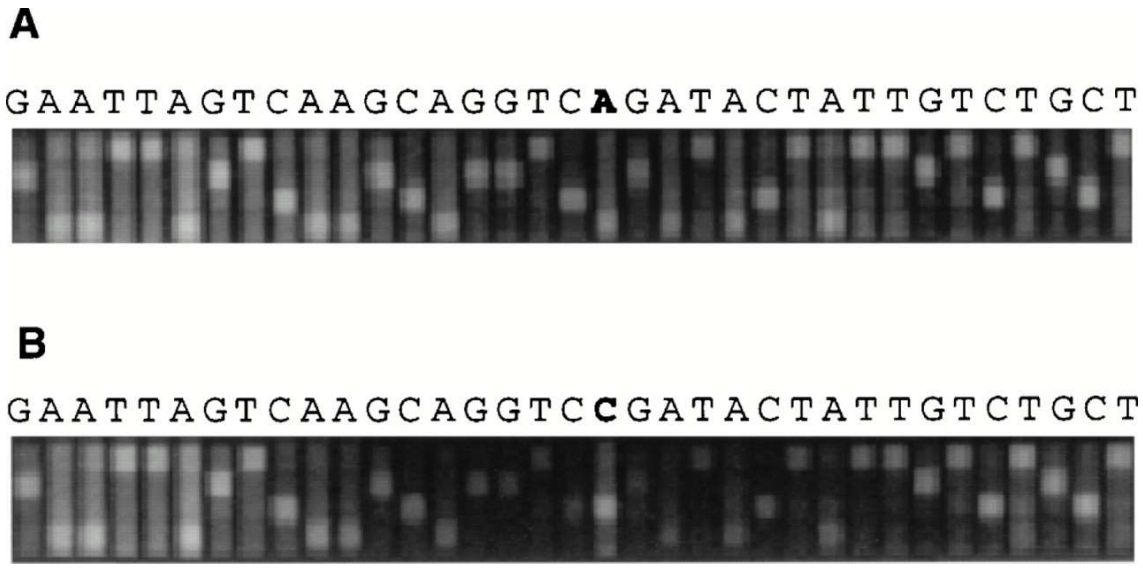


Figure 1.3 A SNP screening chip with two sets of nucleotides that are identical except for a SNP (bold font). In this example, the chromosome shown in panel A carries an 'A' allele, whereas the chromosome shown in in panel B carries a 'C' allele at the same location. Copied from Wang et al. (1998).

SNPs are located throughout the genome, including coding regions and non-coding regions (only 3% of the human genome codes for proteins) (Djebali et al. 2012). Nevertheless, SNPs that are located at coding locations of the genome do not always make changes to the output of a protein. SNPs that *do* cause a change in the amino acid sequence of a protein are named non-synonymous variants. More frequently, SNPs do not cause any change in the amino acid sequence, (named synonymous variants), but can still result in subtle variations to a phenotype, for example due to changing levels of gene expression (Hunt et al. 2009).

Other more complicated genetic variations include insertions and deletions (INDELs), and copy number variants (CNVs). INDELs are polymorphisms in which a section of DNA has either been inserted or deleted, ranging from a single nucleotide to a larger block of hundreds of nucleotide base pairs (Mullaney et al. 2010). Very large insertions or deletions of 1,000 to 400,000 base pairs are called structural variants or copy number variants (CNV) if they are common in the population (Sharp et al. 2005; McCarroll and Altshuler 2007). INDELS and CNVs are less common than SNPs and are estimated to account for approximately 20% of all variants in the human genome (Frazer et al. 2009).

1.3.2 Monogenic Traits

In monogenic traits or diseases, the phenotype is determined by a single genetic variant, usually having a Mendelian pattern of inheritance. Humans are diploid organisms (having two sets of chromosomes), with the potential for two different alleles at each genomic location, meaning individuals can either be homozygous (have a pair of similar alleles at the genetic point of interest) or heterozygous (have different alleles at the same location). In Mendelian traits, the way that these two alleles interact with one another determines whether the phenotype follows a classical dominant, recessive, or more complex pattern. An example of a monogenic disorder is cystic fibrosis. Individuals having two loss-of-function alleles develop the disease, while those with just one defective copy are phenotypically normal; hence inheritance is said to be recessive (Kerem et al. 1989). With regard to ocular phenotypes, although uncommon, syndromic types of high myopia exist that are inherited through Mendelian patterns (Tang et al. 2008). Mendelian inheritance follows simple rules, as shown in Table 1.1 A Punnett square demonstrating the potential offspring of two heterozygous parents. For a

dominant phenotype, offspring with either one or two copies of the defective 'b' allele will be affected. For a recessive phenotype, only offspring homozygous for the 'b' allele will be affected.

B = Normal allele b = Mutant allele	**B**	**b**
B	**BB**	**B**b
B	**B**b	**bb**

Table 1.1 A Punnett square demonstrating the potential offspring of two heterozygous parents. For a dominant phenotype, offspring with either one or two copies of the defective 'b' allele will be affected. For a recessive phenotype, only offspring homozygous for the 'b' allele will be affected.

1.3.3 Complex and Polygenic traits

Polygenic traits are influenced by a number of different loci (as opposed to the monogenic inheritance of Mendelian disorders). Phenotypes that have non-Mendelian inheritance patterns are called 'complex traits' (Lander and Schork 1994). Most polygenic traits are complex, meaning the phenotype is also influenced by gene-gene interactions, gene-environment interactions, and/or non-genetic factors (Lander and Schork 1994). Other ocular traits such as intra-ocular pressure, are complex, with numerous common variants causing subtle differences in phenotype (Gao et al. 2018). The genetic basis of complex disorders can range from a few key genes having a large effect on the outcome, such as in eye colour (Walsh et al. 2011), compared to traits such as height, in which more than 20,000 variants are thought to play a role (Lello et al. 2018).

1.3.4 The Environment

When discussing genetics, the term 'environment' has a broad meaning, encompassing factors both external to the organism but also within the organism and its constituent cells. For example, both the town someone lives in and the hormones in an individual's body could be categorised as environmental variables (Lobo 2008). As environmental effects may be shared within a population or exclusive to the individual, potential gene-environment interactions may also contribute to the presentation of phenotypes. An example of a gene-environment interaction is skin pigmentation caused by the *MC1R* gene and it's response to UV: the *MC1R* gene shows different levels of expression to

depending on the level of UV in the environment, which is turn affects pigmentation (Orazio et al. 2013).

1.3.5 Genotyping and Imputation

Genotyping - determining the alleles present at a specific locus in the genome - can be done directly or indirectly (by imputation). A genotyping 'array' or 'chip' is commonly used to genotype individuals for hundreds of thousands of genetic variants, including SNPs & INDELs (Rabbee and Speed 2006). However, these several hundreds of thousands of directly genotyped SNPs only constitute a small minority of polymorphic sites in the human genome, and will miss most new mutations (LaFramboise 2009). Because of this, whole genome sequencing has been proposed as an alternative method of obtaining genotype data (Kingsmore 2015). However this is an expensive method. Hence, currently, genotyping arrays are the mainstay for genotyping in large-scale studies.

Because SNP arrays only extract a small proportion of known genetic variants, imputation is used to 'gap-fill' the variants not directly assessed (Howie et al. 2012). Using knowledge about genetic variants that *have* been genotyped directly, it is possible to infer the likely alleles that would be present at nearby polymorphic sites that *weren't* genotyped (Howie et al. 2012). Imputation operates by considering 'haplotypes'. A haplotype is a group of closely situated alleles that are commonly inherited together (The International HapMap Consortium 2005). This non-random inheritance can be used to predict unknown nearby SNPs through imputation.

Imputation is performed using reference panels, which are needed to infer which alleles are commonly found together in different populations. An example is the international HapMap consortium reference panel (The International HapMap Consortium 2007). Software such as IMPUTE (including all its updated versions) can be used to match haplotypes and impute non-genotyped SNPs through the use of these reference panels (Bycroft et al. 2018).

Reference panels are developed from genetic data taken from hundreds or thousands of people; the 1000 genomes projects included the full genotyped data from 2,504 individuals (The Genomes Project et al. 2015). This has been done across multiple

populations with different ethnicities and genetic ancestries, with the ability to determine which alleles are more commonly inherited by different populations.

As imputation relies on a reference panel, it performs poorly for alleles that occur more rarely, i.e. variants with low MAF. This can lead to incomplete matching between haplotypes of the individuals in the analysis and the reference panel. IMPUTE software (Wiliams et al. 2012) provides a measure of imputation accuracy, allowing investigators to exclude variants that may be of poor imputation quality and confidence, saving any analyses performed from spurious associations that may occur. However, this diminishes the ability to test rare and low frequency variants for association with a trait, which may contribute a significant amount to the phenotype (Young 2019).

1.3.6 Linkage Disequilibrium

As described in the previous section, imputation is made possible by the non-random transmission of alleles, and the existence of haplotypes (Howie et al. 2012). This non-random inheritance of alleles that are physically close together on a chromosome is termed 'linkage disequilibrium' (LD) (Risch and Teng 1997; Terwilliger and Weiss 1998).

If two genetic variants are in LD with one another, this means there is a statistical correlation between the alleles of the two variants. Therefore, if one of these variants shows a significant association with a trait of interest, the variant that is in LD will also demonstrate an association with the trait. This phenomenon leads to difficulty in determining the causal variant, as many variants in high LD with the causal variant would be associated (Goldstein and Weale 2001). Due to sampling variation, the causal variant will not necessarily be the most strongly associated variant.

The relationship of two genetic variants in regard to their LD is often quantified through the r^2 value (squared correlation coefficient) (Devlin and Risch 1995). An r^2 value of 0 indicates no LD and therefore random assortment. The highest r^2 value of 1 indicates complete LD, where alleles at the two loci are always inherited together.

1.3.7 Heritability

When looking at a complex phenotype, it may be desirable to estimate how much of the inter-individual variation in the phenotype is due to environmental factors, and how much is determined by genetic factors (Hill et al. 2008).

For polygenic traits, genetic effects are usually assumed to act additively, i.e. the polymorphic sites act independently of one another, and the number of copies of each allele has a linear effect on the phenotype (Hill et al. 2008). For example, if one group of individuals carries a single copy of a risk allele at a specific locus while another group carries two copies of the risk allele, the former group would have half the estimated phenotypic effect of the latter. Variants with non-additive genetic effects do exist, however, and include variants with very large effects that give rise to the Mendelian dominant and recessive inheritance patterns described in section 1.3.2.

Heritability is the proportion of phenotypic variation that is attributable to genetic effects. It can be expressed in two ways, either narrow-sense heritability (designated as h^2) or broad-sense heritability (designated as H^2) (Visscher et al. 2008b). h^2 is the amount of variation that is attributed to the additive effects of genetic variants, whereas H^2 includes both additive and non-additive effects. Visscher et al. (2008b) have pointed out that heritability is specific to the population measured, and is potentially influenced by shared environmental factors. Its value therefore varies depending on population demographics, and can fluctuate depending on age and ethnicity.

When a trait is said to have a high heritability, this means that the variation found within the phenotype of interest is largely due to genetic variation. For example, height has been identified as a highly heritable trait (Yang et al. 2010; Pickrell et al. 2016; Lello et al. 2018), meaning it is largely genetically determined. Conversely, if a trait has low heritability, this would mean that non-genetic, e.g. environmental exposures, explain most of the phenotypic variance. Overall, heritability provides an indication of the likely success of genetic prediction (Tenesa and Haley 2013). In general, genetic prediction of traits with high heritability will be more accurate (as they are more reliant on the genetic effects which are estimated).

For many years, heritability has been estimated largely through the use of family and twin based studies (Sanfilippo et al. 2010; Sanfilippo et al. 2011). However, recently, new statistical methods have been developed to allow the estimation of heritability in large samples of unrelated individuals (Bulik-Sullivan et al. 2015b). Genome wide complex trait analysis (GCTA) can be used for this task to calculate the 'SNP heritability', which is an estimation of phenotypic variance attributed to a selection of commonly occurring

genetic variants (Visscher et al. 2008b; Yang et al. 2010). GCTA works through estimating the proportion of the trait variance due to additive effects of SNPs, using a genetic relatedness matrix (an array that allows for the estimation of relatedness between individuals; GRM). The pairwise relatedness of individuals in this GRM is fitted as a random effect in a linear mixed model (see Methods 3.1.5) using restricted maximum-likelihood (Yang et al. 2011b). Hence, this method is often referred to as GCTA-GREML (genomic-relatedness-based restricted estimate of maximum-likelihood) (Yang et al. 2016).

As heritability is sample specific, heritability estimates for the same trait measured in different samples can have a large discrepancy. However, there is a common trend seen between twin/family study derived estimates and those attained using common genetic variants: Twin-based heritability estimates are usually higher than other estimates (Sanfilippo et al. 2011). This discrepancy between heritability estimates has been called the "missing heritability" (Hemani et al. 2013; Zhu et al. 2015; Young 2019). Looking at height as an example, twin and sibling studies have estimated a heritability of roughly 80% (Silventoinen et al. 2003; Visscher et al. 2006). However, when using SNP based analyses, the heritability is lower, with reports finding a SNP heritability of 45%-53% (Yang et al. 2010; Rawlik et al. 2016). When using only genome-wide significant SNPs, the heritability estimate was even lower at 10% (Visscher et al. 2012b). This discrepancy, stemming from the missing heritability phenomenon is commonly seen across most complex traits.

There are several proposed explanations for the missing heritability. Firstly, it should be noted that twin and family studies estimate heritability using all genetic variants regardless of their MAF and effect size. By contrast, h^2 due to GWAS variants typically only use common SNPs (MAF >1%). This means a significant number of genetic variants are not included in SNP based heritability estimates as they may be excluded due to MAF or small effect size (Pritchard 2001; Frazer et al. 2009). LD may also cause difficulty in determining which markers should be used for estimation and limit the accuracy of the estimate, reducing the heritability value obtained (Yang et al. 2010).

Secondly, non-additive genetic effects that are not considered in GWAS studies may also add to the heritability. As well as gene-gene and gene-environment interactions, this

can include transgenerational epigenetic effects such as genetic methylation, in which adjacent C and G nucleotides ('CpG' sites) have a methyl group attached to them, rendering a portion of that gene transcriptionally inactive. It is estimated that an average of 80% of the CpG sites in a human genome are methylated (Ziller et al. 2013), but this varies from person to person and can change over a person's lifetime. These methylation patterns are reprogrammed as an embryo develops and are therefore able to change between generations (Ladstätter and Tachibana-Konwalski 2016). However, because they impact on the transcription of genes, transgenerational epigenetic effects will impact the heritability of the trait.

Thirdly, twin studies may have incorrectly assumed a common environment between siblings. Studies investigating heritability in twins assume that both monozygotic (identical) twins and dizygotic (fraternal) twins are brought up and live in the same environments with the same influences and exposures (Joseph 1998; Kim et al. 2015). Therefore, if monozygotic twins show a higher correlation of a trait than dizygotic twins (who share 50% of their alleles), the differences must be due to genetics (Richardson and Norgate 2005). However, this assumption has been heavily disputed (Joseph 1998; Richardson and Norgate 2005), as environmental exposures have been reported to be less similar between dizygotic twins compared to monozygotic twins. Thus, it may be that the heritability estimates for twin studies may be inflated, which may further contribute to the missing heritability.

1.3.8 Genetic Linkage Studies

In linkage analysis, the inheritance of a trait is investigated in a sample of related individuals, such as a family. For example, Figure 1.4 shows a pedigree in which the transmission of the trait can be monitored over multiple generations (Balding 2006). By looking at which individuals in the pedigree have the trait of interest, and comparing their genomes, it may be possible to identify a genetic region that is common between them that may be the cause of the trait (Risch 1990).

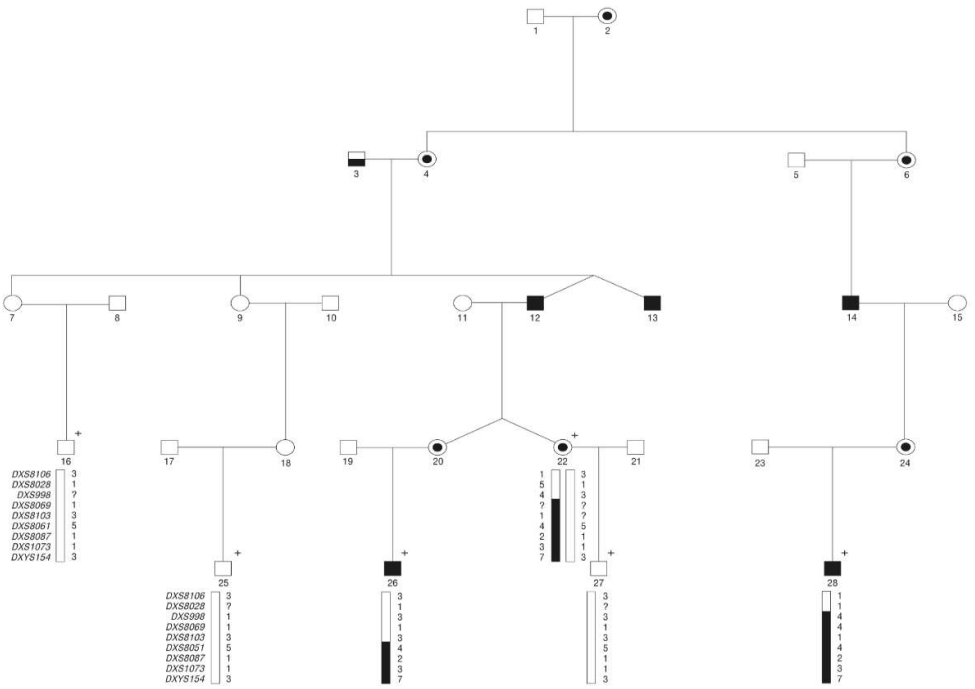


Figure 1.4 Example of a pedigree diagram for Bornholm eye disease, an X chromosome linked high myopia and cone dysfunctional syndrome. Copied from Young et al. (2004). Circles and squares represent males and females, respectively. Affected individuals are highlighted with a solid colour, with carriers shown using partly solid symbols (either a shaded centre circle or half shaded square).

As discussed in the LD section, some variants that are close together and within an 'LD block' are usually inherited together. This means that there are several continuous genetic sequences inherited together, which are separated through recombination (i.e. 'crossing over' during meiosis) (Weiss and Clark 2002). This mechanism allows investigators to perform co-segregation analysis, which is the basis of linkage analysis.

Co-segregation is where a phenotype and alleles at a known polymorphic genetic locus are transmitted together. If co-segregation occurs more than often than expected by chance, it is likely that the casual genetic variant is nearby to the known ('marker') genetic locus. However, it is not possible to narrow co-segregating regions down below the level of LD blocks (which often span across many millions of base pairs and genes), which is one of the major limitations of linkage analysis (Boehnke 1994). Further candidate gene studies are required to investigate genes within the identified region to confirm the causal gene (Lee et al. 2011).

The design of linkage studies means that they are extremely effective when trying to identify rare variants with large effect sizes. This is because a family study would allow

easier identification if the causal variant is uncommon (meaning a smaller pool of people will have the trait and the variant would be easier to identify) and if it has a large effect (meaning that detecting the phenotypic influence is easy) (Hirschhorn and Daly 2005; Sanfilippo et al. 2010). However, a causal locus identified in one family may not demonstrate significant linkage in other families: sometimes loci elsewhere in the genome are responsible for causing a similar phenotype (genetic heterogeneity). Moreover, as linkage analysis relies on segregation of genotypes and phenotypes from parents to offspring, finding families large enough to provide statistically robust results is challenging.

1.3.9 Association Studies

The lowered cost and improved accuracy of genotyping over recent years has facilitated genetic association studies in groups of unrelated individuals. This has led to the adoption of genome wide association studies (GWAS) as a method for identifying causal genetic variants (Visscher et al. 2008a). GWAS analyses test common variants situated throughout the genome for association with a particular phenotype.

GWAS analyses are versatile, in that the phenotype of interest can be either continuous or dichotomous. Conventionally, GWAS variants are assumed to have an additive mode of inheritance (Monir and Zhu 2017; Bonnafous et al. 2018), as described in section 1.3.7. For a continuous phenotype, such additive variants would have an effect size derived from the average phenotype of individuals carrying 0, 1 or 2 copies of the 'risk' allele. For a dichotomous trait, the effect size is quantified as the odds ratio for individuals carrying 0 vs. 1 (or 1 vs. 2) copies of the risk allele.

However, GWAS analyses have limitations. Firstly, as GWAS studies are typically performed using imputed genotype data, spurious results or missed associations may occur due to inaccurate imputation (Howie et al. 2009). To rectify this, GWAS analysts typically remove variants with low imputation quality through quality control processes (Marees et al. 2018).

Similar quality control issues occur when investigating variants with a low occurrence, e.g. MAF < 0.01. Because of their low MAF, there is very limited power to identify true associations correctly (Pritchard 2001). Furthermore, GWAS analyses of dichotomous traits are particularly sensitive to differences in allele frequencies; false associations may

be found if population stratification (uncontrolled population substructure; section 3.1.2) or other population based confounders are not controlled for.

Another issue is the number of statistical tests being performed. Specifically, as numerous variants are being tested one by one, a stringent P-value threshold is required to differentiate between true positive and false positive association signals. A P-value of $5x10^{-8}$ is generally adopted for declaring 'genome-wide significance', because of the expected number of independent polymorphic sites that are tested, and non-random assortment of these due to linkage disequilibrium (Risch and Merikangas 1996; Dudbridge and Gusnanto 2008). Very large sample sizes are therefore required to detect variants with modest effect sizes in order to overcome this problem.

Furthermore, most quantitative or continuous traits are assumed to follow a normal distribution, and if this is criteria is not met it can reduce the power of a GWAS analysis (Goh and Yap 2009). If a trait is not normally distributed, it can be transformed using "inverse rank-based normalisation" before analysis, which has been shown to improve the identification of the causal polymorphisms and improve the summary statistics accuracy (Goh and Yap 2009). However, this method can lead to interpretation difficulties, as the trait dimensions and estimated effects will no longer be on the same scale as the original trait.

1.3.10 Polygenic Risk Scores

A polygenic risk score is a numerical measure used to quantify an individual's genetic predisposition to a specific trait. It is often used to summarise the genetic effects of multiple markers to predict a trait value (or the relative risk of being affected by a disorder, in an intra-group analysis) (Dudbridge 2013). This summary of the estimated genetic contribution to a trait is derived in two steps. Firstly, a GWAS is performed on one sample of individuals, to estimate the effect size associated with each genetic variant tested. The second step, which must be carried out using an independent sample of participants, is to construct the polygenic risk score by taking account of an individual's genotype for each genetic variant and the effect on the phenotype conferred by that variant (Dudbridge 2013). Prediction accuracy can be assessed through a statistical measure such as R^2 or a receiver operating characteristic curve (ROC) analysis, depending on whether the phenotype is continuous or dichotomous. In

theory, if all associated SNPs are included in the analysis model, and the SNP effect sizes estimated very accurately, prediction accuracy should reach a similar level to the 'SNP heritability' (h^2). In practice, most studies have reported much lower accuracies than this. Taking height as an example, although the SNP heritability of height is 0.45-0.53, the best reported polygenic risk score for height had an accuracy of $R^2 = 0.40$ (Lello et al. 2018). It has been proposed that the limitation of this method is largely due to insufficient GWAS sample size, which reduces the accuracy in estimating SNP effect sizes (Visscher et al. 2012a) as well as only using genome wide associated loci, or a limited number of SNPs (Vilhjálmsson et al. 2015).

1.4 Introduction to Myopia Genetics

1.4.1 Heritability

The heritability of refractive error has been estimated using both twin and family based studies as well as SNP-based methods. Twin studies have shown that refractive error has a high heritability, with estimates ranging from approximately 50% to 90% (Lyhne et al. 2001; Wojciechowski et al. 2005; Chen et al. 2007a; Baird et al. 2010; Sanfilippo et al. 2010; Schache and Baird 2012). Additive effects make up the majority of this heritability (Lyhne et al. 2001; Sanfilippo et al. 2010). Family based studies have shown comparable yet reduced estimates of heritability, with a range of 50% to 70% (Chen et al. 2007a; Peet et al. 2007). SNP-based heritability estimates are lower than both of these, as expected, with estimates of 35% - 39% (Guggenheim et al. 2015; Shah et al. 2018).

1.4.2 Linkage Study Discoveries

As linkage studies are best-suited to investigating rare variants of large effect, linkage analysis has mostly been used to study high myopia (at ≤-6.00D) (Farbrother 2003; Zhang et al. 2005). However, there have been studies into refractive error as a continuous trait using this method (Hammond et al. 2004; Klein et al. 2007), as well as low myopia (≤-1.00D) (Stambolian et al. 2004; Chen et al. 2007b). A summary of the 25 myopia loci identified through linkage studies is presented in Table 1.2. The MYP24 and MYP25 loci were identified using a combination of linkage analysis and whole-exome sequencing (a technique that sequences all of the protein coding regions in the genome, known as the 'exome') (Guo et al. 2014; Guo et al. 2015). MYP20 was investigated using a combination of GWAS and linkage methods (Shi et al. 2011b).

Myopia Loci	Chromosome	Myopia Category	Replication Status	References
MYP1	X	High Myopia[7,7]	Replicated	(Guo et al. 2010; Ratnamala et al. 2011)
MYP2	18	High Myopia[7,7]	Replicated	(Young et al. 1998b; Young et al. 2001)
MYP3	12	High Myopia[7,7,5]	Replicated	(Young et al. 1998a; Farbrother et al. 2004; Nurnberg et al. 2008)
MYP5	17	High Myopia[6]	Not Replicated	(Paluru et al. 2003)
MYP6	22	Low Myopia[3,3]	Replicated	(Stambolian et al. 2004; Klein et al. 2007)
MYP7	11	Low Myopia[1]	Not Replicated	(Hammond et al. 2004)
MYP8	3	Low Myopia[1,3]	Replicated	(Hammond et al. 2004; Andrew et al. 2008)
MYP9	4	Low Myopia[1]	Not Replicated	(Hammond et al. 2004)
MYP10	8	Low Myopia[1,3]	Replicated	(Hammond et al. 2004; Stambolian et al. 2005)
MYP11	4	High Myopia[5]	Not Replicated	(Zhang et al. 2005)
MYP12	2	Low/High Myopia[8,2,2]	Replicated	(Paluru et al. 2005; Chen et al. 2007b; Schache et al. 2009)
MYP13	X	High Myopia[7,8]	Replicated	(Zhang et al. 2006; Zhang et al. 2007)
MYP14	1	Low Myopia[4]	Not Replicated	(Wojciechowski et al. 2006)
MYP15	10	High Myopia[8]	Not Replicated	(Nallasamy et al. 2007)
MYP16	5	High Myopia[8]	Not Replicated	(Lam et al. 2008)
MYP17/MYP4	7	Low/High Myopia[4,5]	Replicated	(Ciner et al. 2008; Paget et al. 2008)
MYP18	14	High Myopia[7]	Not Replicated	(Yang et al. 2009)
MYP19	5	High Myopia[7]	Not Replicated	(Ma et al. 2010)
MYP20	13	High Myopia[7]	Not Replicated	(Shi et al. 2011b)
MYP21	1	High Myopia[7,7,7]	Replicated	(Shi et al. 2011a; Tran-Viet et al. 2012; Xiang et al. 2014)
MYP22	4	High Myopia[7]	Not Replicated	(Zhao et al. 2013)
MYP23	4	High Myopia[9,7]	Replicated	(Aldahmesh et al. 2013; Jiang et al. 2014)
MYP24	12	High Myopia[7,7]	Replicated	(Guo et al. 2014; Jiang et al. 2014)
MYP25	5	High Myopia[7]	Not Replicated	(Guo et al. 2015)
MYP26	X	High Myopia[7]	Not Replicated	(Xiao et al. 2016)

Table 1.2 List of myopia loci identified through linkage studies. Thresholds for classifying myopia are indicated with superscript numbers as follows: continuous trait analysis[1], ≤-0.50D[2], ≤-1.00D[3], ≤-3.50D[4], ≤-5.00D[5], ≤-5.50D[6], ≤-6.00D[7], ≤-7.00D[8], ≤-17.00D[9]

1.4.3 GWAS Studies

Initial GWAS studies were performed on small samples of Asian ethnicity, which due to their limited power were not able to find any loci that reached genome-wide significance. Some examples include Nakanishi et al. (2009) who performed a GWAS on 2,741 individuals from Japan, Li et al. (2011) who performed a GWAS on several Singaporean cohorts with a sample of 4,155 individuals, and Shi et al. (2011b) who performed an initial GWAS on 1,088 Han Chinese participants. Larger samples of individuals who had both genetic and refractive data were needed to identify loci at a genome-wide significance level (Hysi et al. 2010; Solouki et al. 2010).

The Consortium for Refractive Error And Myopia (CREAM consortium) performed a GWAS meta-analysis for participants with either European or Asian ancestry (37,382 European and 8,376 Asian participants), from 32 separate population based samples (Verhoeven et al. 2013). They identified 24 novel genome-wide significant loci.

At a similar time, another larger scale study was performed using data from 45,000 individuals from the 23andMe personal genomics company (Kiefer et al. 2013). Participants self-reported if they had been diagnosed with near-sightedness and their age of onset. In this study, 22 loci were identified that were genome-wide significant for age of myopia onset, with 20 shared loci having been identified in the aforementioned CREAM analysis (Verhoeven et al. 2013), and 2 loci being novel. This high level of replication - despite the different phenotypes studied (age of onset of myopia vs. refractive error) - indicated that the two phenotypic measures are closely correlated. Most of the loci identified were near genes related to extracellular matrix structures, photoreceptor functions, eye growth, and neuronal pathways.

More recent GWAS analyses have been performed on even larger samples to further improve statistical power. A GWAS of 191,843 unrelated European participants from 23andMe identified 183 loci for self-reported near-sightedness (Pickrell et al. 2016). This study considered myopia status as a simple binary trait, rather than analysing age of onset as done by Kiefer et al. (2013). Unfortunately, only the top 50 strongest associations were reported in the article published by (Pickrell et al. 2016).

A sample of 160,420 participants were included in a GWAS for refractive error and myopia in a combined analysis by the CREAM consortium and 23andMe. This study

identified 161 independent genome-wide significant loci (Tedja et al. 2018). As with previous CREAM GWAS studies, a combined analysis of European and Asian participants was undertaken (n=44,192 and 11,935 participants, respectively). A second stage of meta-analysis included an additional 104,293 participants of European ancestry from 23andMe. Novel associations near genes responsible for synaptic neurotransmission, anterior segment morphology and light-sensitive signalling cascades were reported.

1.4.4 Genetic Prediction of Refractive Error and Myopia

As refractive error heritability studies have suggested that the genetic variance is largely additive (Lyhne et al. 2001; Sanfilippo et al. 2010), genetic prediction estimates are performed using an additive polygenic risk score approach (section 1.3.10).

The more recent GWAS studies for refractive error and myopia described above have reported the amount of phenotypic variance explained by the genome-wide significant loci discovered, i.e. the prediction accuracy using these loci. Kiefer et al. (2013) estimated the phenotypic variance explained by their 22 genome-wide significant variants to be 2.9%, using a Cox 'survival analysis' model for incident myopia. This was calculated as a *pseudo*-R^2 (similar to a Naglekerke R^2) as their phenotype was a binary outcome. Unfortunately, the genetic prediction analysis was performed in the same 'discovery' dataset as their original GWAS analysis, meaning that the result will be upwardly biased and thus should be interpreted with precaution.

(Verhoeven et al. 2013) found that the 24 loci identified in their GWAS study explained 3.4% of the variance in spherical equivalent (in an independent sample), a modest increase from that found by Kiefer et al. (2013). When considering the capacity to identify individuals as either myopic (≤ -3.00D) or hyperopic (≥ +3.00D), Verhoeven et al.'s 24-variant polygenic risk score had an AUROC of 0.67.

As GWAS sample sizes have increased and new loci for myopia have subsequently been identified, the prediction of refractive error and myopia has improved. When CREAM and 23andMe conducted their GWAS meta-analysis for refractive error, they identified 124 new loci. The 7,307 most-strongly associated variants (those with $P \leq 5x10^{-3}$) explained 7.8% of the variance in refractive error in independent sample, an improvement from the previous estimate and the best reported to date (Tedja et al. 2018). When used to categorize participants as either myopic (≤ -3.00D) or hyperopic (≥

+3.00D), the polygenic risk score derived using these 7,307 variants had an AUROC of 0.77 (Verhoeven et al. 2013).

1.5 Aim of the PhD Project

By better understanding the genetic factors that confer susceptibility to myopia, it may be possible to identify children who are predisposed to become myopic. Moreover, since genetic testing can be performed at any age, genetic prediction of at-risk children could be carried out at an earlier age than is possible using other predictive methods (Zadnik et al. 2015). Such at-risk children may benefit from the interventions described in section 1.2.5 above. To achieve any of this, improvements to the accuracy of genetic prediction of refractive error and myopia are required.

Thus, the primary aim of my PhD project was to leverage information about the genetic contribution to refractive error from the recently released UK Biobank dataset, and then to derive a polygenic risk score based on these new data. The accuracy and clinical utility of the polygenic risk score could then be assessed.

2 Dataset and Participants

The analyses conducted in this thesis were done using two separate datasets collected from the UK. Both of these cohorts are discussed under the relevant subheadings below.

2.1 UK Biobank

UK Biobank was jointly-funded by the Wellcome Trust and the UK Medical Research Council as a prospective study to investigate environmental and genetic elements that contribute to common human diseases (Allen et al. 2014). The study placed a strong emphasis on diseases with a complex aetiology that are projected to increase within the UK population in future years, such as Parkinson's disease and certain cancers (Sudlow et al. 2015). The age range of participants was chosen to be between 40 and 69 years of age to allow longitudinal assessment of these common diseases and their role in premature mortality, and as a compromise between the age at which common illnesses typically occur versus an age range prior to disease onset when risk factor exposure may be important (UK Biobank Team 2007). UK Biobank obtained NHS Research Ethics Committee approval before beginning the study (application reference 11/NW/0382).

2.1.1 Recruitment

Everyone who was registered with the National Health Service and living within 25 miles of a study centre was invited to participate, through the use of posted information sheets. The study comprised of 22 assessment centres spread across England, Scotland, and Wales (Figure 2.1). These were purpose built/renovated centres, usually in town centres that would allow for good transport links and accessibility (Trehearne 2016). 502,682 participants were recruited in total, from 2006 until August 2010.

Figure 2.1 Locations of recruitment centres used in UK Biobank. Adapted from lecture by (Collins 2014), accessed online 27/10/2018, URL: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/06/0940-Collins-UKB-Frontiers-2014-1.pdf.

2.1.2 Self-Reported Medical History

The initial assessment centre visit involved electronic informed consent from all participants. This was performed on touch screen questionnaires, after which demographic information and all medical history, including that of ocular history were collected. A copy of the questionnaire with all questions can be found using the link: https://www.ukbiobank.ac.uk/wpcontent/uploads/2011/06/Touch_screen_questionnaire.pdf?phpMyAdmin=trmKQlYdjjnQIgJ%2CfAzikMhEnx6.

Participants were also given an interview by a registered nurse, along with assessments of cognitive function, physical measurements, and the collection of blood and saliva for genetic study (Sudlow et al. 2015). Although ocular history information was collected for most participants, this was initially limited in scope, and further questions were included on refractive status (self-reported use of spectacles or clear vision for different distances) in the later phases of recruitment. These questions were bundled with an enhanced ophthalmic examination at 6 recruitment centres (see ocular phenotypes below).

In total, participants were asked up to 8 questions regarding their ocular history, including two core questions in the original self-reported questionnaire: “Do you wear

glasses or contact lenses to correct your vision?" and "Do you have any other problems with your eyes or your eyesight?" along with a possible six other follow up questions.

A total of 501,707 people answered if they wore any ocular correction for refraction (89% answered 'yes', 11% 'no', and less than 0.1% 'prefer not to answer'), (UK Biobank Team 2018). If applicable, this was followed with "What age did you first start to wear glasses or contact lenses?" which was answered by 444,542 participants. 16,088 of respondents did not know the age of their first visual correction, while 243 people preferred not to answer.

The enhanced questionnaire introduced during the later recruitment phase included a question on the reason why ocular correction was needed i.e. asking if the participant was myopic, hyperopic, or presbyopic, and which eye the condition affected. A total of 143,561 participants answered this question, with the option of selecting multiple reasons. 64,868 stated they wore spectacles for myopia, 33,415 for hyperopia, 67,615 for presbyopia, and 19,259 for astigmatism. A further 10,995 reported that their correction was for less common conditions such as amblyopia or strabismus, with 2,774 stating that they did not know or preferred not to answer (UK Biobank Team 2018).

2.1.3 Phenotype Information

Six assessment centres in the study (5 in England and 1 in Wales) performed ophthalmic assessments on participants (Cumberland et al. 2015). 117,279 participants (making up 23% of the entire cohort) underwent this enhanced assessment, which included several ocular measurements, including refractive error measured with non-cycloplegic auto-refraction using a Tomey RC 5000 autorefractor/keratometer (Tomey Corp., Nagoya, Japan). The refractive error was measured up to 10 times on each eye, and rated for its reliability from 0 to 9, with any score ≤ 4 considered to be reliable (lower scores were deemed as more reliable). Any poor reliability readings were excluded from the analyses before averaging. These results were then transformed to average mean spherical equivalent, by adding the spherical component and half of the cylindrical component power together. This was then averaged between the two eyes. The equation for this phenotype, Autorefraction Mean Spherical Equivalent (Autorefraction MSE) is shown in Equation 2.1.

$$\text{Autorefraction MSE} = \frac{((R.Sph_1 + 0.5 * R.Cyl_1) + \ldots (R.Sph_n + 0.5 * R.Cyl_n))/n + ((L.Sph_1 + 0.5 * L.Cyl_1) + \ldots (L.Sph_n + 0.5 * L.Cyl_n))/n}{2}$$

Equation 2.1. Where R *and* L *are for the right and left eye, respectively, and n is number of valid measurements taken.*

Individuals who had declared they had any previous eye surgery or certain ocular pathologies were excluded from analysis. This was because a history of complications or surgery may change the refractive error measurement from what the patient would have naturally developed, introducing error in the phenotype e.g. refraction can change due to IOL surgery or cataract development. Answering 'yes' to any of the following resulted in exclusion (Plotnikov et al. 2019):

- Previous injury or trauma resulting in loss of vision
- Any serious eye problems
- Previous cataract surgery or glaucoma surgery e.g. lens extraction or trabeculectomy
- Any other previous ocular surgery in the last 4 weeks
- Refractive laser eye surgery
- Corneal graft surgery
- Self-reported cataracts or retinal detachment

This resulted in 42,617 participants being removed from analyses in which autorefraction-measured refractive error was the phenotype of interest.

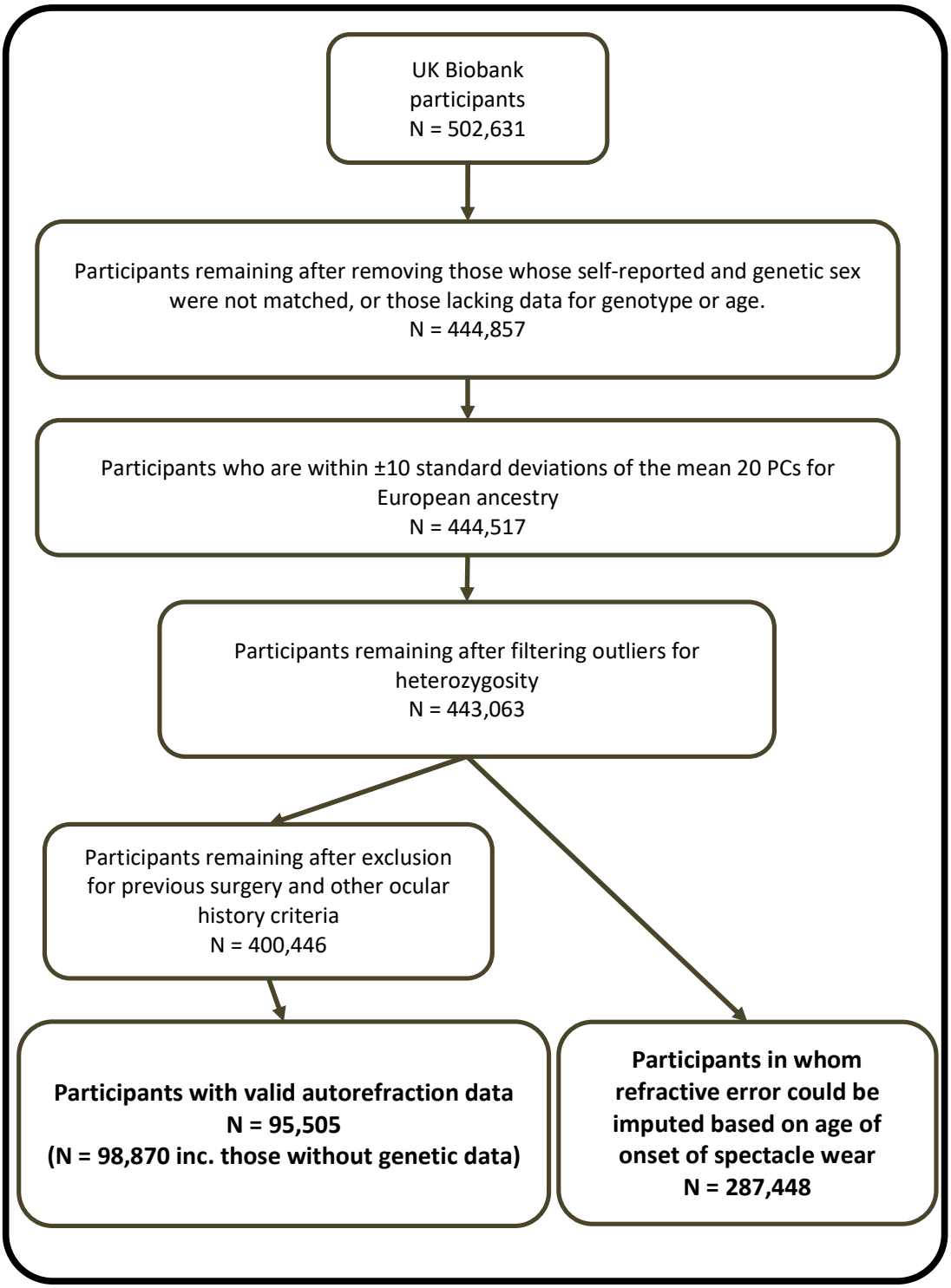
2.1.4 Genotype Information

DNA samples from participants were genotyped by means of genotyping arrays. UK Biobank used two different genotype arrays; the UK Biobank Lung Exome Variant Evaluation (BiLEVE) Axiom array or the UK Biobank Axiom array (Bycroft et al. 2018). A subset of 49,950 individuals were genotyped on the BiLEVE array, which had some distinct variants chosen because of their suspected involvement in lung function and related diseases. All other participants were genotyped with the Axiom array, specifically designed for UK Biobank. Overall these genotype arrays shared 825,927 (95%) of markers between them. Genetic data were released in two phases, of which

the second wave included data for 488,377 participants. Data from this second phase were used in all analyses.

From the ~800,000 directly-genotyped markers on the genotyping array, Bycroft et al. (2018) imputed genotypes at ~87 million loci in the genome using IMPUTE4 (see section 1.3.5 in Chapter 1). This was done using a reference panel created from the UK10K and Haplotype Reference Consortium (HRC), while the 1000 Genomes Phase 3 panel was used for phasing (The 1000 Genomes Project Consortium et al., 2015; The UK10K consortium et al. 2015; Loh et al., 2016). Genetic data were available for 443,063 participants after undergoing quality control filters for heterozygosity, a mismatch between self-reported and genetic sex, and non-European ancestry. A full flowchart of the participants available from UK Biobank is shown in Figure 2.2.

Figure 2.2 (overleaf) Flowchart indicating the steps taken to filter UK Biobank participants to the groups used in several analyses in this thesis, which included participants with European ancestry. The number of participants with refractive data who had self-reported European ancestry but did not have genetic data has also been stated for the analysis in chapter 6. For this subgroup, those with self-reported 'white' ancestry who had all covariate data and did not have any phenotypic exclusion criteria were included.


UK Biobank participants N = 502,631
Participants remaining after removing those whose self-reported and genetic sex were not matched, or those lacking data for genotype or age. N = 444,857
Participants who are within ±10 standard deviations of the mean 20 PCs for European ancestry N = 444,517
Participants remaining after filtering outliers for heterozygosity N = 443,063
Participants remaining after exclusion for previous surgery and other ocular history criteria N = 400,446
Participants with valid autorefraction data N = 95,505 (N = 98,870 inc. those without genetic data)
Participants in whom refractive error could be imputed based on age of onset of spectacle wear N = 287,448

2.1.5 Cohort Limitations

Although UK Biobank is a valuable resource for investigating complex traits, it does have limitations that should be considered, particularly for ocular traits. For example, as already stated above, not all participants underwent refractive error measurement. Thus, although UK Biobank has genotype information on over 480,000 individuals, only 117,279 of them had refractive error data. After data quality control filters and exclusion criteria have been applied, the resulting dataset comprises fewer participants than other studies reported in the literature (Pickrell et al. 2016; Tedja et al. 2018). Therefore, in my analyses, consideration was given into how to utilize the participants in the cohort who did not undergo autorefraction in order to increase the effective sample size for GWAS analyses (see section 4.1).

The self-reported difficulty in distance/near vision that participants were asked was only ascertained for approximately 143,000 participants, most of whom also underwent the ophthalmic assessment. Therefore, any prediction or estimation of refractive error using this source of information in participants without autorefraction measurements would be limited due to the small sample size (only 37,000 participants without refractive error information were asked these questions).

The demographics of the UK Biobank sample do not accurately represent the UK population. The aim of the UK Biobank project was to study common complex diseases, which tend to be more common in older individuals. Therefore, the target population age range of the study (40-70 years old) meant that children and young adults were not represented (Sudlow et al. 2015). Moreover, the study has an upward bias in its age range, with the greater majority of participants being 55 years of age or older. Therefore, UK Biobank is limited in its applicability to address research questions focussing on adolescent myopia development; for example, myopia reported by older individuals may be due to crystalline lens changes. However, many such cases of cataract-associated myopia would have been excluded using the criteria specified in Section 2.1.3 (Plotnikov et al. 2019). In addition, as UK Biobank participants would have spent their childhood in an era before smartphones and mobile devices were available, the adolescent environmental risk factor profile of the UK Biobank cohort may not reflect that of children growing up today.

The UK Biobank study has recruited more female than male participants on average (Fry et al. 2017). The socioeconomic status of UK Biobank participants was also higher than the UK population, with individuals more likely to own their own home and live in an area with a lower deprivation index (Fry et al. 2017). Furthermore, there is a 'healthy participant selection bias' overall within the cohort, with participants being leaner and less likely to smoke and demonstrating a lower mortality rate than the UK general population. The most important consequence of the lack of representation with the UK population is the potential for biased associations arising due to collider bias (Fry et al. 2017; Munafò et al. 2018).

The majority (95%) of participants in UK Biobank are of "white" ethnicity (i.e. European ancestry), which is representative of the 2001 census population (Fry et al. 2017). However, this meant that to control for population stratification (section 3.1.2), restriction to the largest homogenous ethnic sample (Europeans) would be required for the majority of the analyses performed, as failing to do so would present higher numbers of false positives due to confounding (McClellan and King 2010). Moreover, further difficulty arises because many participants are related; this again can lead to bias in genetic association studies (Thomson and McWhirter 2017). Therefore, when trying to maximise statistical power of my GWAS studies, consideration was also given regarding overcoming potential bias due to relatedness of individuals.

2.2 ALSPAC cohort/Children of the 90's

The Avon Longitudinal Study of Parents and Children (ALSPAC), also known as 'The Children of the 90's study' is a population-based, collaborative study working with the European Longitudinal Study of Parents and Children (ELSPAC), which in turn was designed due to the World Health Organisation (WHO) identifying the need for longitudinal studies in different environments to understand modifiable elements to children's development (Boyd et al. 2013). Original funding came from WHO Europe, with other sources of funding obtained from the early methodology and questionnaire pilots. The data currently available for the children are from birth up to 18 years of age, with the prospect of having lifelong follow-up results. Some information was collected by means of self-reported questionnaires and parental questionnaires. Physical measures were also taken during visits to the ALSPAC research clinic. Ethical approval

for all aspects of the study were obtained through the ALSPAC Ethics and Law Committee, along with Local Research Ethics Committees.

2.2.1 Recruitment

The ALSPAC team recruited 14,541 expectant mothers (as the initial remit of the study was to evaluated the influences of the environment during pregnancy) who were due to give birth between April 1 1991 and December 31 1992 from the former county of Avon in South West England (Boyd et al. 2013). This included the city of Bristol, other nearby smaller towns, and surrounding rural areas, but excluded the city of Bath. Informed consent was obtained from all participants, along with the children's agreement per visit.

Expectant mothers were recruited via media advertisements and visiting community locations in the catchment area. If any interest was shown, or further details requested, an information pack was sent, with an 'opt out' system. Twelve months after the end of the recruitment phase, 13,988 of the babies were alive and included in the study.

Recruitment had two additional phases. One was an attempt to recruit parents who had initially opted-out to re-join when their children were aged 7 years old, another when the children were aged 8 years old. Finally, 15,247 children were included in the study.

2.2.2 Phenotype Collection

Measurements of the children's refractive error were obtained using non-cycloplegic autorefraction with a Canon R50 (Canon USA, Inc., Lake Success, NY, USA) at the ages of 7, 10, 11, 12, and 15 years old (Shah et al. 2017). At least 3 readings were taken at each visit, which were averaged. Participation was subject to attrition; not all participants attended appointments at all ages, with fewer returning for appointments at older ages. Some participants in the cohort did not attend for any refractive error or ocular assessments; 9,401 children presented at least one vision assessment clinic visit, meaning that 5846 (38%) had no refractive error data (Boyd et al. 2013). The average mean spherical equivalent refractive error at each clinic visit was calculated using equation 1 above (as per the UK Biobank cohort).

Parents accompanying ALSPAC children attending the research facility when their child was 7 years old were also invited to sit for non-cycloplegic autorefraction if there was

sufficient time remaining. Overall, 1,516 ALSPAC parents (all of whom were mothers, aged 24 and over) took up the invitation to sit for autorefraction, which was performed with the same Canon R50 instrument.

2.2.3 Genotype Collection

Researchers collected blood samples from both children and mothers, extracting DNA from immortalised lymphocytes. DNA from the ALSPAC mothers was analysed using the Illumina 660 W-quad chip (Fraser et al. 2013), whereas DNA from children was genotyped using the Illumina HumanHap550 quad chip genotyping platform (Taylor et al. 2016). This was performed for all children and mothers who gave consent. This resulted in over 10,000 children and mothers with genetic data (Fraser et al. 2013)

As per the UK Biobank cohort, the majority of participants in ALSPAC are of European ancestry. After excluding related individuals (e.g. twins), those with poor quality genotype data, and participants who withdrew their consent, 7,981 children, and 1,516 mothers remained in our sample who had available phenotypic data and were genotyped. After combining information from markers common to both genotyping arrays, genotypes for 477,482 variants were available. Imputation was performed (by the ALSPAC research team) with IMPUTE2 (Howie et al. 2009) against the first phase of the 1000 Genomes reference panel.

2.2.4 Cohort Limitations and Exclusion Criteria

As mentioned in the phenotype section above, not all participants attended clinic visits at all available ages. There were 7,852, 7,310, 6,575, 6,582, and 4,899 children with refractive error measurements available at the ages of 7, 10, 11, 12, and 15, respectively. Furthermore, the numbers of participants at each age reduced to 5,634, 5,369, 4891, 4895, and 3728 at ages 7, 10, 11, 12, and 15, respectively, for those with genotype data. Only 2,048 individuals with genetic data had refractive error measurements available for all 5 ages, with a general trend towards participants attending during early years of childhood and dropping out as they got older.

As previously mentioned, the mothers of the ALSPAC children were not the primary focus of the original study, hence parental refractive error was measured only if time permitted. Repeated measures on each eye were not performed as consistently with

the ALSPAC mothers as with the ALSPAC children i.e. a different number of readings for each eye was done on the mothers (sometimes perhaps only one reading per eye). This may have resulted in less precise estimates of the true refractive error.

ALSPAC children who attended for all visits tended to have a more myopic refractive error than those that did not, with an average MSE of +0.20D for children with 4 or less visits at the age of 15, against average MSE of +0.18D at the age of 7 for children that attended at all visits. However, this is not significant and unlikely to bias the result (*t* test, P = 0.56).

Another important point to note is that the final autorefraction measurement in ALSPAC children was performed at the age of 15, when the phenotype may not have reached the level obtained by adulthood. This will mean that some participants may have been categorised as non-myopic, when in fact they will go on to develop myopia at an older age (Williams et al. 2013; Fan et al. 2016a), presenting a form of survivor bias (in which many people who have avoided becoming myopic before this age would be categorised as non-myopic hereafter). This mis-categorisation could lead to a measurement error, and thus limit the accuracy of genetic prediction.

3 General Methods

This chapter describes the commonly-used methods employed in the experimental chapters of the thesis. When software applications are discussed, I describe the settings applied, along with a brief background explaining the theoretical basis underlying the tests performed.

3.1 Data Preparation for GWAS analysis

In order to perform a GWAS, data on the phenotype of interest, any covariates, and genotype are required. These data need to be organised so that information pertaining to each individual is matched and that it meets a certain level of quality. Therefore, before executing any GWAS analysis, quality control and exclusion criteria need to be applied to the data.

3.1.1 Data File Formatting

For GWAS software packages to be able to read the input data, the data must be stored using a standard format. Historically, a popular format has been the PLINK text format (Purcell 2007), which comprises two file types: a 'ped' file containing information about the individuals, such as their phenotype and genotypes, and a 'map' file which includes information about the genetic markers and their genomic location. More recent versions of PLINK have switched to a three file format that uses compact binary coding for the genotype data: 'bed' binary files contain the participant ID and genotype data, 'fam' files contain individual and phenotype information, and 'bim' files contain details of the alleles and physical location of the markers (Purcell et al. 2007; Chang et al. 2015; Loh et al. 2018). As reading large text files can take a long time, the binary file format offers improved speed (Marees et al. 2018). However, further advances have been made, with new BGEN binary files created to handle even larger datasets (Band and Marchini 2018), such as UK Biobank (Bycroft et al. 2018), in which genotytpe data can be handled with a faster processing speed due to greater data compression. BGEN format files can be accessed with most commonly used GWAS software (Chang et al. 2015; Loh et al. 2015b). Thus, BGEN format files were used in Chapters 5 and 6 when analysing data from UK Biobank and PLINK binary format files were used in Chapters 4, 7, and 8 when analysing ALSPAC data.

3.1.2 Principle Component Analysis

Different ancestral populations have undergone genetic drift and thus can have different allele frequencies, which may lead to the case where a rare variant in one ancestry is relatively common in another (Byun et al. 2017). Using these differences in allele frequency and common genotype patterns it is possible to determine population structure and ancestry (Linck and Battey 2019). However, these differences between sub-populations can cause confounding bias (known as "population stratification") when conducting an association study. As GWAS testing is sensitive to these allele frequency differences, testing in a diverse, non-homogenous sample may lead to spurious associations on account of this population stratification (Price et al. 2006). For example, if a case-control investigation was performed for brown iris colour in a sample of Asian and Caucasian participants together, it would lead to false-positive associations as brown eyes are more common in Asian populations. Specifically, markers with a higher allele frequency in either Asians or Europeans would exhibit a spurious (non-causal) association with eye colour.

Originally, 'genomic control' was proposed to limit the inflation of P-values that would occur because of the high number of false positive associations from population stratification (Reich and Goldstein 2001; Devlin et al. 2004). However, this method has no ability to distinguish between true associations or false positives; genomic control corrects for stratification uniformly, leading to a loss of power. This is discussed later in section 3.2.1.

Because of this potential confounder, the use of principal component analysis (PCA) has been proposed as an alternative or additional solution (Reich and Goldstein 2001; Price et al. 2006). PCA works by determining patterns and trends in multi-dimensional data to infer structured genetic variation (Patterson et al. 2006). Then these identified patterns can be used to convert the correlated data into a list of linearly uncorrelated variables, transforming the data to have a reduced number of dimensions, which means that greater population trends (such as common allele frequency differences) can be controlled for (Price et al. 2006). PCA has the added advantage that it can reduce the number of correlated dimensions by any desired amount; i.e. the researcher can control for several different dimensions rather than one, which has the added benefit of allowing partitioning of multiple ancestries (Byun et al. 2017). Therefore PCA is

commonly used in GWAS analyses as a method to identify participants' ancestral backgrounds and to control for sub-population trends (Bycroft et al. 2018; Marees et al. 2018). An example of a PCA analysis for race is shown in Figure 3.1.

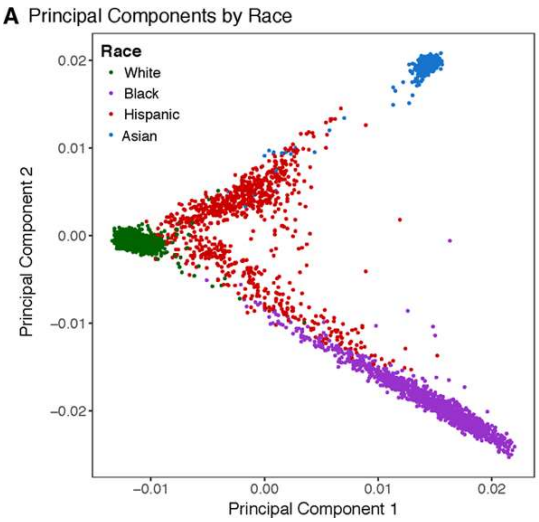


Figure 3.1. Example of a graphical representation of principle component analysis differentiating principle components 1 and 2 for different ethnicities. Taken from Khera Amit et al. (2019).

For the investigations in this thesis, the PCs calculated by Bycroft et al. (2018) were used. The top 20 PCs were included as quantitative covariates in all GWAS analyses in order to reduce false positive associations due to population stratification. PCs were also utilized to identify a group of UK Biobank participants with European ancestry. The mean and standard deviation values for the first 20 PCs for individuals who self-reported their ethnicity as 'White' were calculated (444,517 participants). Then, participants whose PCs were within ±10 standard deviations of the mean for all 20 of the top PCs were considered as a homogenous group of European ancestry that could be analysed together in a GWAS analysis (N = 443,063) (please see section 2.1.3).

3.1.3 Covariates

For all participants in UK Biobank, the age when the participant attended was recorded, including when they had their non-cycloplegic autorefraction performed. This age value was used as a quantitative covariate in all GWAS analyses. The sex of each participant was also used as a binary categorical co-variable, and any sex mismatched individuals (a difference found between self-reported and genotype-inferred sex) were excluded from analysis on account of potentially poor quality genotype data.

As described in section 2.1.4, 2 different genotyping arrays were used for the collection of genetic information (either the BiLEVE or UK Biobank Axiom array) (Howie et al. 2009; Bycroft et al. 2018). To ensure the use of different genotyping arrays did not induce any spurious associations from differences in imputation, genotype array type was used as a categorical covariate. Any individuals who had missing data for the phenotypic trait of interest, or had missing covariate information were excluded from the GWAS analysis.

3.1.4 Additional GWAS Filters

Both genetic and phenotypic information was filtered to ensure that it was suitable for GWAS analysis. The filters chosen were in accordance with those commonly used in GWAS publications (Marees et al. 2018).

The maximum genetic data missing per individual and maximum missing per SNP were set at 5%, and 2% respectively, meaning that any person or SNP with missing data greater than these percentages would be excluded. Any SNP that deviated from Hardy Weinberg Equilibrium (HWE) was excluded. Usually, violations of HWE indicate genotyping error, or non-randomised mating/evolutionary selection (Knapp et al. 1995). The threshold for this filter was set at $P < 1x10^{-6}$, which has been advised for quantitative traits (Marees et al. 2018). Participants with extremely high or low heterozygosity were also excluded, specifically those outside of ±4 standard deviations from the mean. This measure of genetic diversity would allow those with an atypical level of genetic variation to be excluded from the analysis (extremely high heterozygosity is suggestive of two DNA samples accidentally being mixed together; extremely low heterozygosity is indicative of a poor quality DNA sample that produces numerous incorrect genotype calls).

3.1.5 BOLT Software for GWAS Analysis

BOLT-REML LMM version 2.3.1 (BOLT) was used for conducting all GWAS analyses in this thesis. BOLT is a commonly used software package able to run GWAS analyses of both continuous and categorical phenotypes while accounting for relatedness, both familial and cryptic (Loh et al. 2015b; Loh et al. 2018).

BOLT does this by using a linear mixed model in the association testing for single polymorphic sites in a GWAS. This contrasts with many other GWAS software packages,

such as PLINK (Purcell 2007; Chang et al. 2015), which use simple linear or logistic regression models rather than a mixed model.

Linear mixed models (LMM) have similar features to a simple linear regression, in which there are fixed effect estimations used in the model. However, linear mixed model can account for random effects within the data, which can be used to account for underlying population stratification and relatedness in the test sample. BOLT creates a genetic relatedness matrix (GRM), in which the degree of relatedness (kinship) between pairs of individuals is estimated through the similarity of their genotypes. BOLT conducts association tests using a 'leave one chromosome out' (LOCO) process, in which variants situated on the same chromosome as the test SNP are not assessed in the random effects of the linear mixed model during analysis (Loh et al. 2015b). This is done so that the test SNP is not included twice in the model, in both the fixed effects when tested for association and in random effects through relatedness. If the LOCO method was not used this would lead to an over-fitting of the model due to proximal contamination, and reduce the test power (Listgarten et al. 2012).

This use of a LMM is a crucial improvement over traditional linear regression analyses used in GWAS, as previously being unable to account for relatedness led to reduced power in detection of associations and an increased number of false positive association signals and inflated P-values (Yang et al. 2014). The implementation of a linear mixed model to control for relatedness enables the inclusion of related individuals without introducing confounding bias, increasing the sample size available for GWAS testing and thus improving statistical power (Korte and Farlow 2013).

To create the genetic relatedness matrix in BOLT, a set of variants genotyped with a high degree of certainty are recommended (Loh et al. 2015b). These are termed 'high confidence' variants. For my analyses, 890,000 high confidence variants were selected, namely those with 'rs' prefixes, genotyped in at least 99% of individuals, which had a MAF of >0.05. These SNPs were LD pruned (see Section 3.2.5 for details) using a 50 marker wide window, with one marker from any pair of markers removed if they had an LD $r^2 > 0.1$. The windows were advanced in 5 marker steps.

BOLT can implement two different models when performing GWAS analyses, an infinitesimal model, and a non-infinitesimal model (Loh et al. 2015b). The infinitesimal

model assumes that all variants in the analysis are causal, with each having a small effect the size of which is normally distributed. The non-infinitesimal model is more complicated, implementing a Bayesian approach that assumes effect sizes of genetic variants do not follow a single normal distribution as per infinitesimal models, and that a small number of variants with large effects are present, with the remaining variants having smaller effect sizes. These different probability distributions are calculated by BOLT, based on the input data, and incorporated into the linear mixed model. Should there be an improvement with the use of non-infinitesimal models, BOLT would signal this in the final output file. As there was no such improvement for any of the GWAS tests performed, all results were taken from the infinitesimal model analysis.

3.2 Additional Genetic Analyses

3.2.1 Genomic Inflation Factor

For all GWAS analyses performed, a genomic inflation factor (λ_{gc}) was calculated as the median observed χ^2 statistic divided by the expected median χ^2 statistic. This was used alongside quantile-quantile (QQ) plots to assess for possible genomic inflation. Should the λ_{gc} be greater than 1, this may indicate some population stratification bias that would need to be addressed. As mentioned above, adjusting GWAS summary statistics based on λ_{gc} has been argued to be overly stringent, as it does not account for potential polygenicity of the trait, in which a trait with many variants causing small effects leads to a higher λ_{gc} value (Bulik-Sullivan et al. 2015b). Therefore, although λ_{gc} was calculated to explore potential inflation in the GWAS tests conducted, it was interpreted in the context of other software (LDSC; described below) that allowed the assessment of polygenicity.

3.2.2 LD Score Regression: LD Score Regression Intercept

To assess whether λ_{gc} had a value greater than 1 due to polygenicity, the LD Score Regression (LDSC) Intercept (λ_{LDSC}) was calculated to compare against λ_{gc}. The LDSC intercept is an estimation of the inflation of association test statistics due to population structure or genotyping errors (Bulik-Sullivan et al. 2015b). LDSC uses reference LD information from the HapMap3 reference panel for individuals with European ancestry (The International HapMap et al. 2010; Bulik-Sullivan et al. 2015b).

Initially an LD score for all variants included in the GWAS is calculated. The LD score is calculated using a regression analysis that examines the relationship between GWAS summary statistics for SNPs and linkage disequilibrium scores for those SNPs, computing the sum of the pairwise r^2 values between the variant and all other variants within 1 centimorgan (cM) (on average, 1 cM corresponds to approximately $7.5x10^5$ base pairs) (Lodish H 2004). By doing this, it is possible to calculate SNP heritability estimates and genetic correlation (see below) estimates between pairs of traits using only GWAS summary statistics (Bulik-Sullivan et al. 2015a).

The LD score intercept was calculated for all GWAS summary statistics, to test for any potential P-value inflation due to population stratification.

3.2.3 LD Score Regression: Genetic Correlation

LD score software also allows the calculation of genetic correlations between pairs of traits using GWAS summary statistics. LDSC takes advantage of LD patterns in that GWAS effect size estimates for a SNP would encompass the effects of other SNPs in LD with that SNP (Bulik-Sullivan et al. 2015b). If a trait is polygenic, SNPS with higher LD scores would, on average, produce higher χ^2 statistics relative to SNPs with lower LD scores (Yang et al. 2011a). LDSC works by calculating and controlling for LD patterns in the genome, and then assessing the genetic effects for two specified traits (Bulik-Sullivan et al. 2015a). This can then calculate the genetic correlation between these two traits; a value is given between -1 and 1, in which a value close to 1 or -1 would indicate that the genetic influences of two traits are identical, and a value of 0 would indicate complete independence in the traits genetic effects. All genetic correlations reported in this thesis were calculated using LDSC software.

3.2.4 Multi-Trait Analysis of GWAS (MTAG)

Multi-Trait Analysis of GWAS (MTAG) software (Turley et al. 2018) was used as the primary statistical method for combining summary statistics from GWAS analyses of different traits should they demonstrate a genetic correlation. The standard approach for combining GWAS summary statistics is inverse-variance weighted meta-analysis, through software such as METAL (Willer et al. 2010), which has been used in previous GWAS analyses for refractive error (Verhoeven et al. 2013; He et al. 2014; Tedja et al. 2018). However, MTAG has been argued to perform as well, or even better than

standard meta-analysis since, (i) it is computationally as fast as standard meta-analysis methods (by assuming that the variance–covariance matrix of effect sizes across the traits is the same), (ii) it permits the use of overlapping data samples when used for different phenotypes, and (iii) when used to combine information across multiple traits, it outputs regression coefficients that are specific for the trait-of-interest. For example, if the genetic correlation between Trait A and Trait B is 0.5, then conventional meta-analysis would yield regression coefficients corresponding to a hybrid of these two, e.g. two GWAS results for one SNP with the effect sizes of +0.4 and +0.2 would (assuming equal standard errors) give an output of +0.3 in the meta-analysis. However, with MTAG, a separate regression coefficient output would be given for Trait A and for Trait B. MTAG also offers the benefit that output values can be transformed into the same units as the trait of interest, maintaining its translational nature when merging correlated traits measured using different units.

The fundamental idea underpinning MTAG is that when GWAS estimates from different traits are correlated, the effect estimates for each trait can be improved upon by incorporating information contained in the GWAS estimates for the other traits (Turley et al. 2018). Traits can demonstrate correlation for two main overarching reasons: either from a true genetic correlation, or the estimation error of variant effects may be correlated between traits. This will usually be due to phenotypes of the two traits demonstrating correlation or underlying biases found between the SNP effect estimates, from sources such as population stratification or relatedness. MTAG uses these two sources of correlation to increase the statistical power of effect estimates for the trait of interest (Turley et al. 2018). All MTAG analyses were performed using the default settings, with the assumption of an infinitesimal SNP effects model (Turley et al. 2018).

3.2.5 LD Control: LDpred Software

When calculating a genetic risk score for genetic prediction, the linkage disequilibrium (please see Section 1.3.6) between SNPs should be taken into account. This is because the presence of highly correlated SNPs being used together can lead to several SNPs essentially displaying the same information for the identified loci (i.e. the same information about a significant association with that section of the genome is given by several SNPs). Moreover, non-causal positive associations in GWAS studies make it

challenging to identify causative SNPs; a high correlation between SNPs (i.e. high LD), means that when trying to detect an association with a trait, several candidate SNPs will be identified. This can lead to inaccuracies in downstream analyses such as genetic risk score creation and trait risk estimation due to the use of non-independent predictors (Liu et al. 2013; Vilhjálmsson et al. 2015).

A commonly reported approach to account for LD is 'clumping and thresholding' (Rydzanicz et al. 2011). Clumping is the process of taking index variants (i.e. variants identified as being 'significantly' associated with the trait-of-interest according to a pre-specified p-value threshold), and grouping ('clumping') the index variant together with all nearby variants in a region of a pre-determined genomic length. The non-index clumped variants are then excluded from the downstream analysis. In LD-based clumping, instead of clumping all variants in the region of pre-specified length around each index variant, only variants with a pre-specified level of LD with the index variant are clumped together. Again, the non-index clumped variants are excluded from the downstream analysis.

A typical LD-based clumping threshold would be an $r^2 = 0.1$ and a typical region-based clumping distance would be 250 kb, i.e. 250,000 base pairs either side of the index SNP (Vilhjalmsson et al., 2015). After clumping and thresholding has been performed, SNPs in high or moderate LD with the index SNPs and within the same genomic region are removed from the analysis. However, the use of clumping and thresholding has limitations. The region size for clumping is chosen arbitrarily, and dismisses informative markers that may be nearby one another during the selection process, thereby limiting the obtainable phenotypic variance explained (Flister et al. 2013). The P value threshold chosen for defining index SNPs is also arbitrary.

Another method of filtering genetic markers in LD is 'LD pruning' (Calus and Vandenplas 2018). This is done by identifying a SNP and computing its LD (r^2) with nearby SNPs in a pairwise fashion. Should 2 SNPs have LD higher than a certain threshold, one of the SNPs is removed (usually the one with the lower MAF). This process continues iteratively until there are no SNPs in LD remaining. Unlike clumping and thresholding, LD pruning can remove strongly associated SNPs (those with low P-values in a GWAS), and may lead to the output of only a few SNPs, or areas of the genome that have no representative SNP

at all (i.e. there is no grouping on relative distance of variants). This in turn may lead to a reduced number of variants being used in calculating a genetic risk score, thus reducing predictive accuracy. Thus, it has been suggested that LD-based clumping is superior to traditional LD pruning (Marees et al. 2018).

An even more effective methodology proposed to account for LD is to down-weight variants that are in LD in the same way they would be in a regression model that fitted all variants simultaneously rather than one-by-one. A widely-used implementation of this approach is the LDpred software package (Vilhjálmsson et al. 2015). LDpred uses a Bayesian approach in calculating posterior mean effect sizes for all GWAS markers, based on LD patterns that are present in a reference panel with the same ancestry as the investigated sample.

LDpred analysis has three steps. Firstly, genotype data from the reference panel is used to calculate LD (r^2) values for pairs of markers. Secondly, these values are used to adjust the effect size weights of summary statistics, incorporating the LD structure from the first step, using Gibbs sampling. A Gibbs sampler is a computational process using a Markov chain (in which the probability of an event depends on the state of a previous event or known value), which is used to calculate a posterior probability distribution. With regard to LDpred, this step involves the use of pairwise LD correlations obtained from an ancestry-match reference panel to infer the distribution of the genetic markers and the relative likelihood of pairs of alleles being inherited together. After running several iterations, the posterior probability distribution can be assessed to determine the likely genotype-phenotype relationship in the sample conditional on the genotype at other marker locations. The final step involves the application of these new SNP weights to an independent validation dataset (see the Polygenic Risk Score section 1.3.10 below).

This approach has been shown to improve the accuracy of genetic prediction compared to clumping and thresholding methods (Vilhjalmsson et al., 2015). Because of this, LDpred software (version 1.0.6) was used in chapters 7 and 8 with the aim of improving the accuracy of the genetic risk scores.

3.2.6 Polygenic Risk Score Estimation

All genetic risk score calculations were performed using the PLINK software '--SCORE' function (Purcell 2007). This included the third (validation) step of LDpred analysis. Genetic risk scores were calculated using a formula in which the effect size of each SNP was weighted according to the degree of association in a GWAS analysis (i.e. the weightings corresponded to GWAS regression coefficients) and multiplied by the number of risk alleles present. All risk scores assumed an additive model as explained in Section 1.3.7. Thus, the risk score for each participant was calculated as the sum of the contribution from each locus; no interaction effects were modelled. The formula for the genetic risk score was:

$$\textit{Predicted refractive error of an individual} = (E_1 * X_1) + (E_2 * X_2) + (E_3 * X_3) + \ldots (E_n * X_n)$$

Equation 3.1 Equation for the genetically predicted refractive error for any given individuals. E_i is the regression coefficient (also known as "beta coefficient" or "effect size") for variant i*, and X_i is the number of risk alleles of variant* i *carried by the participant of interest.*

3.3 Statistical Analyses

All other statistical analyses, such as logistic regression, linear regression, linear mixed models, survival analysis, and receiver operating characteristic curve (ROC) analysis were performed using R (version 3.5.0). Creation of tables and correlative data was done using a mixture of R and Microsoft Excel software (R scripts for analyses where applicable are in the Appendices, Chapter 1). Quality control filtering of genotype and phenotype data was performed using custom-written BASH scripts. GWAS analyses, use of LDpred, MTAG and LDscore were performed using the Cardiff University ARCCA RAVEN Supercomputing Cluster.

4 Prediction of Refractive Error in Children Using Either Genetic Risk Scores or Number of Myopic Parents

4.1 Introduction

The CREAM consortium have previously identified 149 variants that were genome wide significantly associated with refractive error in a CREAM meta-analysis and that demonstrated some evidence of replication ($P < 0.05$) in an independent sample (Tedja et al., 2018). In this experiment, these 149 variants were used in combination to create a polygenic risk score for predicting refractive error in children.

A person's number of myopic parents has been used for many years by researchers to estimate the risk of developing myopia (Hui et al. 1995; Wu and Edwards 1999; Mutti et al. 2002; Jones-Jordan et al. 2010), as described in Section 1.2.4. In this respect, the number of myopic parents (0, 1, or 2) has been proposed as a predictor variable that captures familial factors associated with the inheritance of myopia. Moreover, many of the above studies made the assumption that number of myopic parents primarily captures the genetic risk of myopia, thereby ignoring environmental influences that may also be associated with parental myopia.

In this chapter, number of myopic parents was compared to the genetic risk score in order to assess the extent to which parental myopia reflects inherited genetic risk and/or other sources of influence. (The latter sources of influence would include 'genetic nurture' (Zhang et al. 2015a; Richmond et al. 2017), i.e. non-inherited genetic risk arising from environmental effects associated with non-transmitted parental genotypes, as well as other environmental factors associated with parental myopia). Specifically, the hypothesis tested in this study was that using the genetic risk score would improve the prediction accuracy of refractive error in children beyond that of knowing their number of myopic parents. As well as examining the prediction of *refractive error* per se, a secondary aim was to test the hypothesis that the genetic risk score would also enhance the prediction of the *incidence of myopia* better than knowing the number of myopic parents alone.

4.2 Methods

The methods are split into four sections: The determination of the best genetic risk score model, linear regression models for refractive error prediction accuracy, refractive error prediction trajectories, and the prediction of incident myopia. A summary of the participants used in the different sections of this chapter are shown in Figure 4.1.

4.2.1 Study Participants

The children studied were participants in the Avon Longitudinal Study of Parents and Children (ALSPAC; see Section 2.2). The children were invited to attend the research clinics for a host of measures on an annual basis, although not all children attended at all years. Ocular measurements were taken at 5 different age visits, when the participants were aged 7, 10, 11, 12, and 15 years old (Williams et al. 2008b). Refractive error was measured using non-cyclopegic autorefraction. The average mean spherical error was calculated as described in Section 2.1.3.

Genetic data were obtained through analysis of DNA extracted from blood samples. Approximately 10,000 ALSPAC children were genotyped (Boyd et al. 2013). There were 7,981 children with full genome-wide genotype information remaining after excluding data that failed quality control assessments, participants that withdrew consent, participants that were related, and participants of non-European ancestry (in order to avoid issues relating to population stratification).

Parental myopia was inferred using self-reported answers to an item on a questionnaire completed, separately, by the child's mother (or guardian) and by their father (or guardian) when the mother was pregnant. Each parent was asked to complete the following question: "How would you rate your sight without glasses?" The options given for response were: "always very good", "I can't see clearly at distance", "I can't see clearly close up", and "I can't see much at all." Reponses were classified into being myopic or not, as a binary variable. This classification was optimised by comparing the answers given by parents whose refractive error was measured. Thus, the responses were categorised using the following methodology (Shah et al. 2017): parents whose responses for both eyes were, "I can't see clearly at a distance" or "I can't see much at all" or a combination of these two responses were classed as being myopic. Parents with both eyes categorized as "always very good" or "I can't see clearly close up" or a

combination of these two responses were classed as being non-myopic. Any other combination of responses resulted in the classification being set as “missing”.

4.2.2 Selection of Genetic Variants

As explained in the introduction, the 149 lead variants from the CREAM and 23andMe meta-analysis (Tedja et al., 2018) that demonstrated evidence of replication in an independent sample were included in the genetic risk score creation. These variants were selected because they were from the largest GWAS for refractive error that had been published, and since the ethnicity of the CREAM and 23ndMe GWAS sample (European) matched the ethnicity of the majority of the ALSPAC participants.

4.2.3 Genetic Risk Modelling

The PLINK software “--SCORE” function was used to generate genetic risk scores (as described in Section 3.2.6). Genetic risk scores were calculated using two different methods.

The first genetic risk score model was an “allele score” (also termed a “raw genetic risk score”). In this model, the effect size is not used in the calculation and all variants are weighted equally in the creation of the genetic risk score i.e. the effect size is 1 for all variants, and therefore only the number of variants is counted and summed together to give the risk profile.

The second model was a weighted genetic risk score, as described in Section 3.2.6, which should theoretically provide greater precision than the first model. The variant weights (beta coefficients) for this model were obtained from the GWAS for refractive error in UK Biobank participants. Hence, the genetic risk score “effect size” in this model would be expressed in dioptres. However, for ease of interpretation, the scores were standardised to have a mean of zero and a standard deviation of 1, i.e. converted to Z-scores.

Each child in the sample was assigned a genetic risk score (using both models) based on the genotypes for the 149 genetic variants. Linear regression models were created to identify how well these genetic risk scores explained the observed variance in refractive error (see below). The genetic risk score model (i.e. model #1 or #2, as described above) that gave the best accuracy in predicting refractive error was then compared against the

predictor variable “number of myopic parents” (coded as a categorical variable: 0, 1 or 2). A combined model comprising of the genetic risk score as well as the number of myopic parents was assessed to test whether the combined model improved the accuracy of refractive error prediction above that of either predictor alone.

4.2.4 Refractive Error Linear Model Prediction

Multivariable linear regression models were created using number of myopic parents and/or genetic risk score as predictor variables for refractive error (in Dioptres). Number of myopic parents was coded as a categorical variable with three levels (0, 1, or 2 myopic parents), with zero myopic parents used as the reference category. Genetic risk score was coded as a continuous variable (model 1: unweighted allele score; model 2: Z-score for weighted allele score). The better-performing of these two genetic risk score models was used for further analysis to compare against number of myopic parents. These prediction models were applied using all participants who had their refractive error measured at either 7 or at 15 years old. In order to allow a comparison of predictive performance for children of different ages (i.e. age 7 vs. 15 years old) a sample of participants who had refractive data available at both ages was analysed to allow for a direct comparison.

4.2.5 Estimation of Refractive Error Trajectory

The ALSPAC children’s longitudinal data gives the opportunity to model the development of refractive error. Linear mixed models were used to calculate refractive trajectories in a similar method to that done by Fan et al. (2016a). This was done for number of myopic parents, genetic risk score, and a combined model, for all children who had attended for at least 3 visits to measure refractive error (as not all children had attended for all ages). Fixed effects included in the model were: sex, age-at-visit (to the first, second, third, and fourth polynomial order), and the predictor variable(s) of interest (i.e. number of myopic parents, genetic risk score, or both). Individual-level random effects included the refractive error of each child in the baseline age 7 visit (random intercept), and the trajectory (random slope) of refractive error with age.

The results for these models were plotted to display the estimated refractive trajectory. This was done to demonstrate the trajectories for children between the ages of 7 – 15 years old with: 0, 1, or 2 myopic parents, a high, average or low genetic risk score, and

a combined model that categorised children into one of nine groups, determined by their number of myopic parents and their genetic risk category. An “average” genetic risk was defined as having a standardised genetic risk score of zero, and “low” and “high” genetic risk were defined as a genetic risk score of 1 standard deviation lower or higher than the mean value, respectively.

4.2.6 Prediction of Myopia Incidence

In order to investigate if number of myopic parents, genetic risk score, or a combined model was more accurate at predicting *incident myopia*, survival analysis was conducted using Cox proportional hazards models (Breslow 1975; Guggenheim et al. 2012).

Participants with a known number of myopic parents, a genetic risk score, and who had at least one auto-refraction measurement were included in these analyses (for more information please see Figure 4.1). Because ALSPAC children had their refractive error measured using non-cycloplegic autorefraction (Boyd et al. 2013), and this has been shown to give an average -0.25D measurement error compared to that of cycloplegic autorefraction in ALSPAC (Williams et al. 2008a; Northstone et al. 2013), participants were classified as myopic if their mean spherical equivalent was ≤ -1.00D. Age of onset of myopia was defined as the age of the child at the first clinic visit at which their refractive error was below the myopia threshold. For children who were classified as non-myopic at all of their clinic visits, their age at their last clinic visit was used as the right-censored time to event.

Three models were created: a model for number of myopic parents (0, 1 or 2), a model for genetic risk score categorised as low, average, or high risk, and a model with both predictors. The genetic risk score category for these last two models was defined as above in the LMM methods (resulting in 3 categories in model 2, and 9 categories in model 3). All models included sex as a predictor variable.

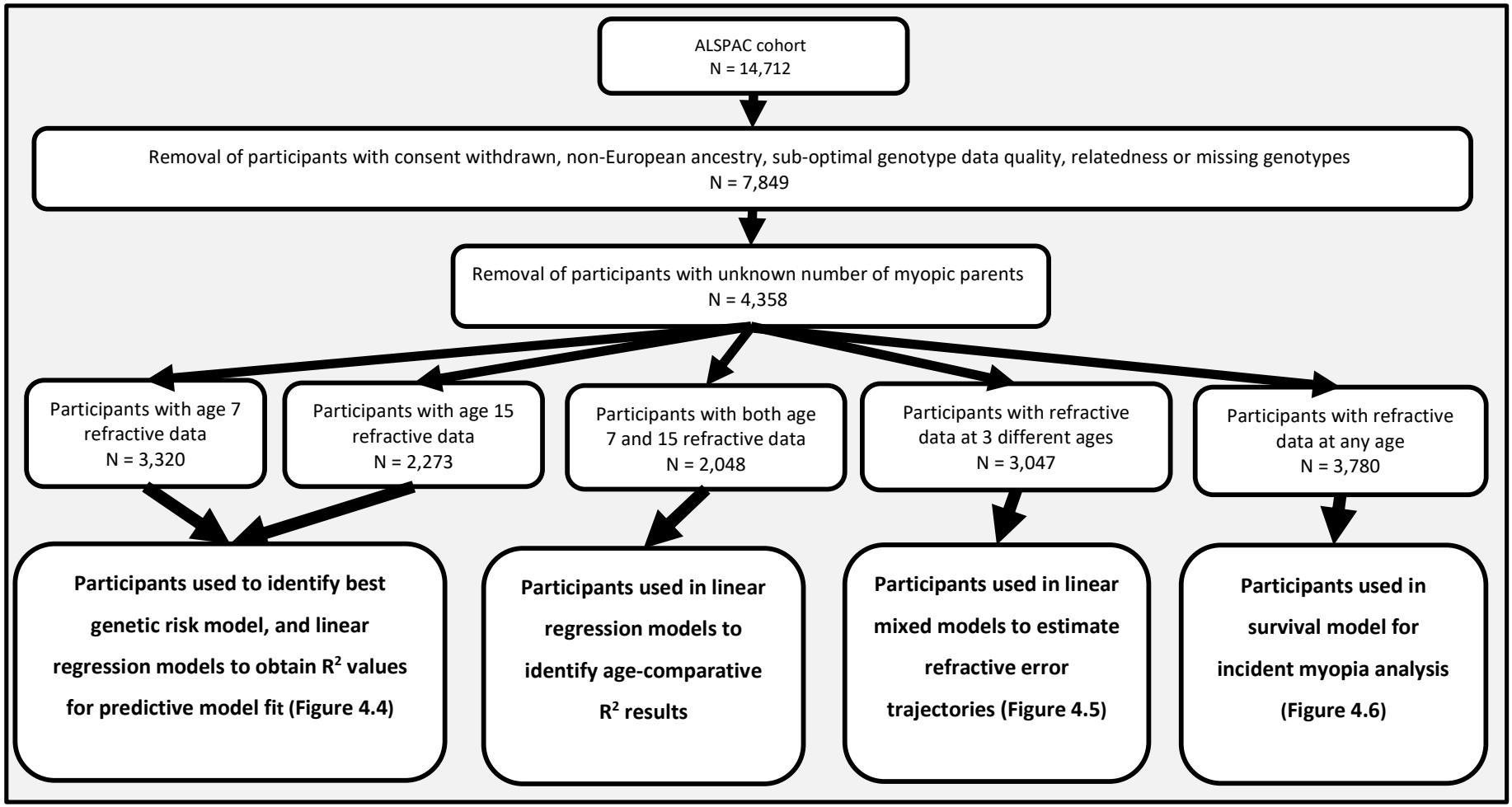


Figure 4.1 Flowchart demonstrating participant selection (adapted from Ghorbani Mojarrad et al. (2018)).

4.3 Results

Population demographics for the participants used in the analyses at ages 7 and 15 are shown in Table 4.1. There were a total of 3,350 and 2,273 children with a known number of myopic parents, genotype data, and who had autorefraction measurements at age 7 or 15 years of age, respectively. An overlapping sample of 2,048 participants had refractive data for both ages 7 and 15 years old.

Variable	**Age 7 years sample (N=3,320)**	**Age 15 years sample (N=2,273)**	**Age 7 and 15 sample (N=2,048)**	
			Age 7	**Age 15**
Age (mean ± SD)	7.47 ± 0.17 yrs	15.41 ± 0.27 yrs	7.45 ± 0.14 yrs	15.41 ± 0.26 yrs
Percentage male (N)	51% (1,680)	47% (1,071)	48% (978)	
Refractive error (mean ± SD)	+0.17 ± 0.81 D	-0.43 ± 1.24 D	+0.16 ± 0.79 D	-0.43 ± 1.19 D
Percentage myopic (N)	2.1% (71)	16.2% (369)	2.3% (48)	16.1% (329)
Number of myopic parents				
Zero (%)	1341 (41%)	886 (39%)	783 (38%)	
One (%)	1535 (46%)	1052 (46%)	963 (47%)	
Two (%)	444 (13%)	335 (15%)	302 (15%)	

Table 4.1. Demographics of the samples used in the linear regression prediction models. Samples from the first two rows were also used in the identification of the best genetic risk score model.

4.3.1 Genetic Risk Score as a Predictor Variable

Two different genetic risk score models were evaluated, an unweighted model (model #1) and a weighted model (model #2). The predictive performance of these two models is displayed in Table 4.2. The distribution of the number of risk alleles carried by the participants is shown in Figure 4.2A, while Figure 4.2B demonstrates the distribution with standardised Z score values.

The average number of risk alleles carried by the sample was 121 (with a range of 95 to 151). The maximum possible would be 298 (149 risk variants x 2 copies of each allele). The weighted genetic risk score model (model #2) demonstrated a normal distribution (Shapiro-Wilks normality test W = 0.94, P = 0.1), whereas the number of risk alleles (model #1) did not (Shapiro-Wilks normality test W = 0.91, P = 0.04). By definition, the

mean and standard deviation of the standardised model #2 were approximately 0 and 1, respectively.

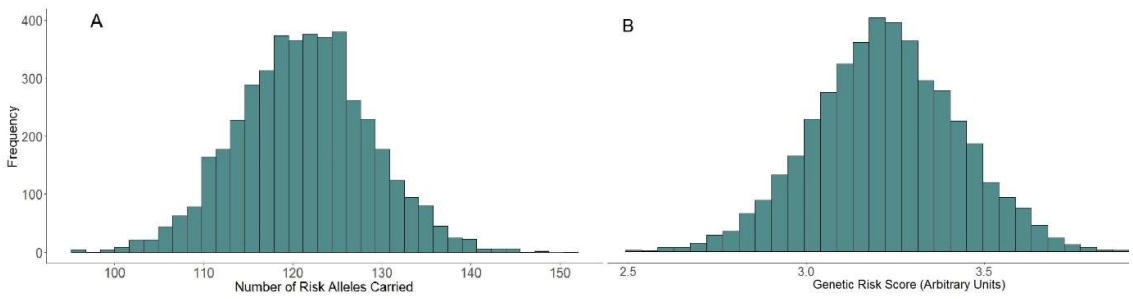


Figure 4.2 Histograms demonstrating the different distributions of A: number of risk alleles carried, B: a transformed standardised genetic risk Z score.

The performance of the two models in predicting refractive error in children at 7 and 15 years old is displayed in Table 4.2. The results show that the R^2 when using the weighted Z score model was slightly higher at age 15 (model fit P = 0.06 and P = 0.05 for age 7 and 15, respectively). Therefore, weighted genetic risk scores were used for all subsequent analyses in this chapter.

Age	Model #1 (Unweighted allele score)	Model #2 (Standardised, weighted Z score)
7 Years Old	0.0087 (0.001-0.016)	0.0114 (0.004-0.019)
15 Years Old	0.0222 (0.009-0.035)	0.0264 (0.013-0.039)

Table 4.2. Performance in predicting refractive error for 2 different genetic risk score models. Values indicate the variance explained; R^2 (95% confidence intervals).

4.3.2 Number of Myopic Parents as a Predictor Variable

There were subtle yet significant differences found between number of myopic parents and the distribution of genetic risk scores. In comparison to participants with no myopic parents, the genetic risk score was 0.15 standard deviation units higher in those with one myopic parent, and 0.36 standard deviation units higher in children with 2 myopic parents (both $P < 0.00001$). The distribution of genetic risk scores in children with different numbers of myopic parents is illustrated in Figure 4.3.

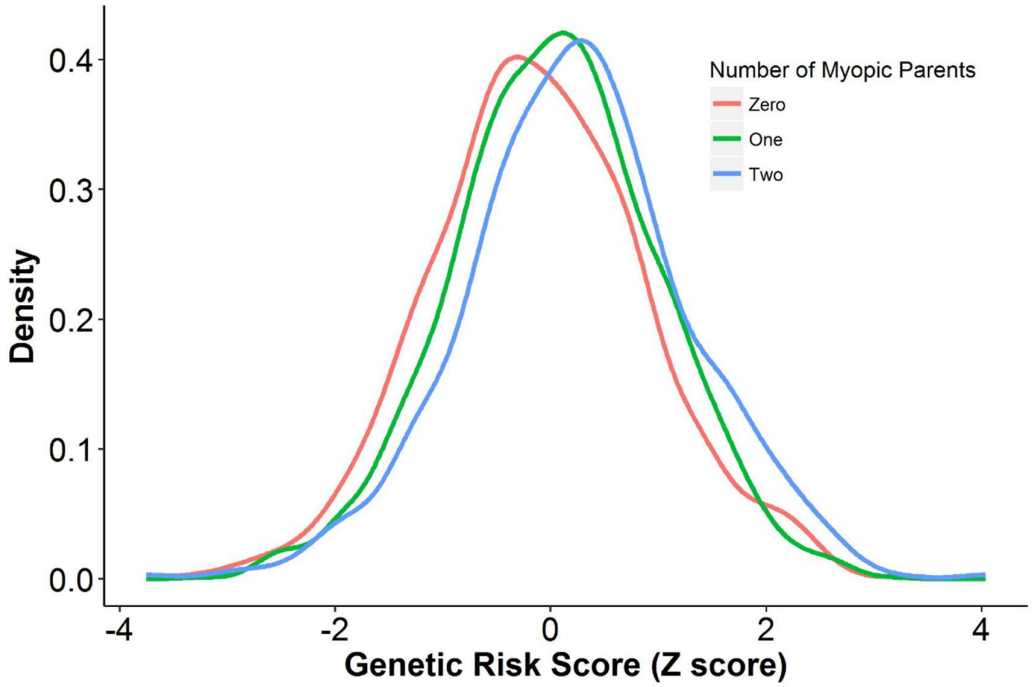


Figure 4.3. Density plot displaying the distribution of genetic risk scores (Z-scores) for participants with 0, 1, and 2 myopic parents. Note that the sample sizes of these groups were not equal; there were 1,859, 1,946, and 553 participants with 0, 1, and 2 myopic parents, respectively. Adapted from Ghorbani Mojarrad et al. (2018).

4.3.3 Refractive Error Linear Model Prediction Results

4.3.3.1 Number of Myopic Parents

For ALSPAC participants aged 7 years old, the variance in refractive error explained (R^2) by the number of myopic parents was 3.0% (95% CI 1.8-4.1%, $P < 2.2x10^{-16}$). At the age of 15, the variance explained increased to 4.8% (95% CI 3.0-6.5%, $P < 2.2x10^{-16}$).

4.3.3.2 Genetic Risk Scores

As described above, the weighted genetic risk score was also weakly predictive of refractive error at both ages. At age 7, the variation in refractive error explained was 1.1% (95% CI 0.04-1.9%, $P = 4.6x10^{-10}$). This increased to 2.6% (95% CI 1.3-3.9%, $P = 5.1x10^{-15}$) at the age of 15 years.

4.3.3.3 Combined Genetic Risk Z Score and Number of Myopic Parents

Combining the predictor variables together improved predictive performance at both ages. At 7 years, the R^2 was 3.7% (95% CI 2.5-5.0%, $P < 2.2x10^{-16}$) and at 15 years, 7.0% (95% CI 5.0-9.0%, $P < 2.2x10^{-16}$). Adding an interaction term for both variables in the model showed no evidence for an interaction between genetic risk score and number

of myopic parents in the combined model at either age ($P > 0.05$). An improvement was found when comparing the combined models to the model with number of myopic parents alone (likelihood ratio test: $P = 1.1 \times 10^{-7}$ and 2.4×10^{-12}, at ages 7 and 15, respectively). A summary of these results can be seen in Table 4.3 and Figure 4.4.

4.3.3.4 Prediction Comparison at Ages of 7 and 15 Years of Age

Prediction comparisons between children aged 7 and 15 years were carried out for the subset of 2,048 participants with information available at both ages (see Figure 4.1). When only the number of myopic parents was used as the predictor variable, the R^2 increased from 2.8% (95% CI 1.4-4.2%) when measured at age 7 years, to 4.6% (95% CI 2.8-6.4%) at age 15 years. When using solely the genetic risk score, the R^2 increased from 0.7% (95% CI 0.0-1.4%) at age 7 years to 2.0% (95% CI 0.8-3.2%) at age 15 years. For prediction using a combined model, the R^2 increased from 3.3% (95% CI 1.8-4.8%) at age 7 years to 6.1% (95% CI 4.1-8.0%, $P < 2.2 \times 10^{-16}$) at age 15 years. Although the R^2 values were higher at the age of 15 than at the age of 7, the overlapping confidence intervals mean that it is difficult to draw any firm conclusions as to whether the difference between prediction comparisons was significant.

	Model A Number of myopic parents			**Model B** Genetic risk score			**Model C** Combined analysis			Model 1 vs. Model 3
	R^2	95% CI	P-value	R^2	95% CI	P-value	R^2	95% CI	P-value	P-value
Children aged 7 years (N=3,320)	0.030	0.018-0.041	$<2.2x10^{-16}$	0.011	0.004-0.019	$4.6x10^{-10}$	0.037	0.024-0.050	$<2.2x10^{-16}$	$1.10x10^{-7}$
Children aged 15 years (N=2,273)	0.048	0.030-0.065	$<2.2x10^{-16}$	0.026	0.013-0.039	$5.1x10^{-15}$	0.070	0.050-0.090	$<2.2x10^{-16}$	$3.93x10^{-12}$

Table 4.3. Accuracy (R^2) in predicting refractive error using linear regression models with predictor variables: number of myopic parents, genetic risk Z score, and a combined analysis model. The model A vs. model C significance was tested using a likelihood ratio test.

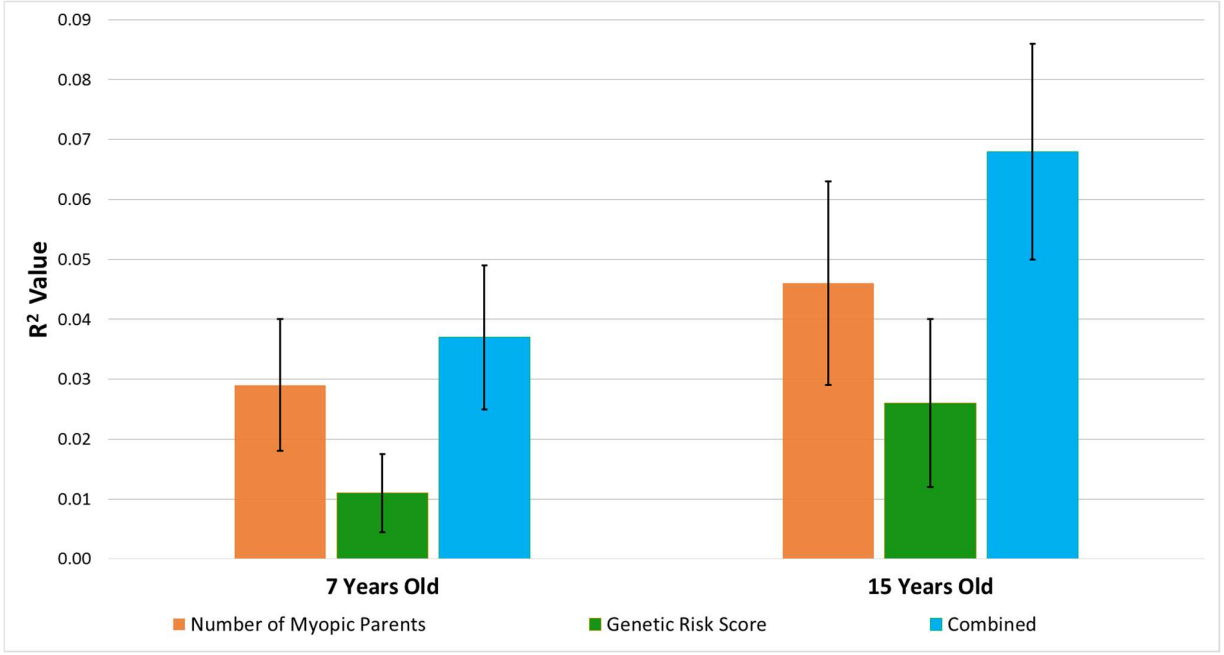


Figure 4.4. Bar chart illustrating the accuracy of predicting refractive error using number of myopic parents, genetic risk score, and a combined model, in children aged 7 or 15 years old. Error bars are 95% confidence intervals (adapted from Ghorbani Mojarrad et al., (2018)).

4.3.4 Linear Mixed Model Refractive Error Trajectories

3,047 participants were included in the linear mixed model analyses after excluding participants without genotype data or information about their number of myopic parents, and those who had attended fewer than 3 visits at different ages (Figure 4.1).

Figure 4.5 displays the trajectories predicted using the best-fit models for number of myopic parents, genetic risk score and a combined model, respectively. All models indicated a progression towards a more myopic or negative refractive error with age across for all risk categories. This shift was less observable in the lower risk categories for both genetic risk score and number of myopic parents.

A widening in the distribution of refractive error with age was evident in all three analyses, with the number of myopic parents showing a larger variation between individuals with different levels of the risk factor (i.e. 0, 1, or 2 myopic parents) at the age of 15 years old compared to the high vs. low genetic risk score. The graphical data suggest that children with 2 myopic parents have a higher degree of myopia at the age of 15 than those in the high genetic risk score category.

Figure 4.5C illustrates the trajectory of participants' refractive error using the combined predictor model. This model appears to stratify participants into a wider refractive error range than the other two models, with the lowest risk found in children with zero myopic parents and a low genetic risk score, whilst the highest risk was identified in participants with two myopic parents and a high genetic risk. For the combined model, there was evidence for a 3-way interaction between genetic risk score, number of myopic parents, and age of visit ($P = 4.3x10^{-3}$).

4.3.5 Prediction of Incident Myopia

Both a higher number of myopic parents and a higher genetic risk score were predictive of a higher rate of incident myopia. The survival curves from the Cox proportional hazard models are shown in Figure 4.6. Comparing panels A and B suggested that the increased risk of myopia conferred by having two vs. zero myopic parents was larger compared to having a high vs. low genetic risk score. The results from the combined model for incident myopia suggested that participants with two myopic parents and a high genetic risk score had the highest myopia incidence, while children with no myopic parents and

a low genetic risk had the lowest incidence rate. Inclusion of genetic risk score in the survival model (panel C) improved the model fit compared to using the number of myopic parents alone ($P = 4.02 \times 10^{-9}$). A test for an interaction between the number of myopic parents and the genetic risk score in the combined model showed no evidence of such an interaction ($P > 0.05$).

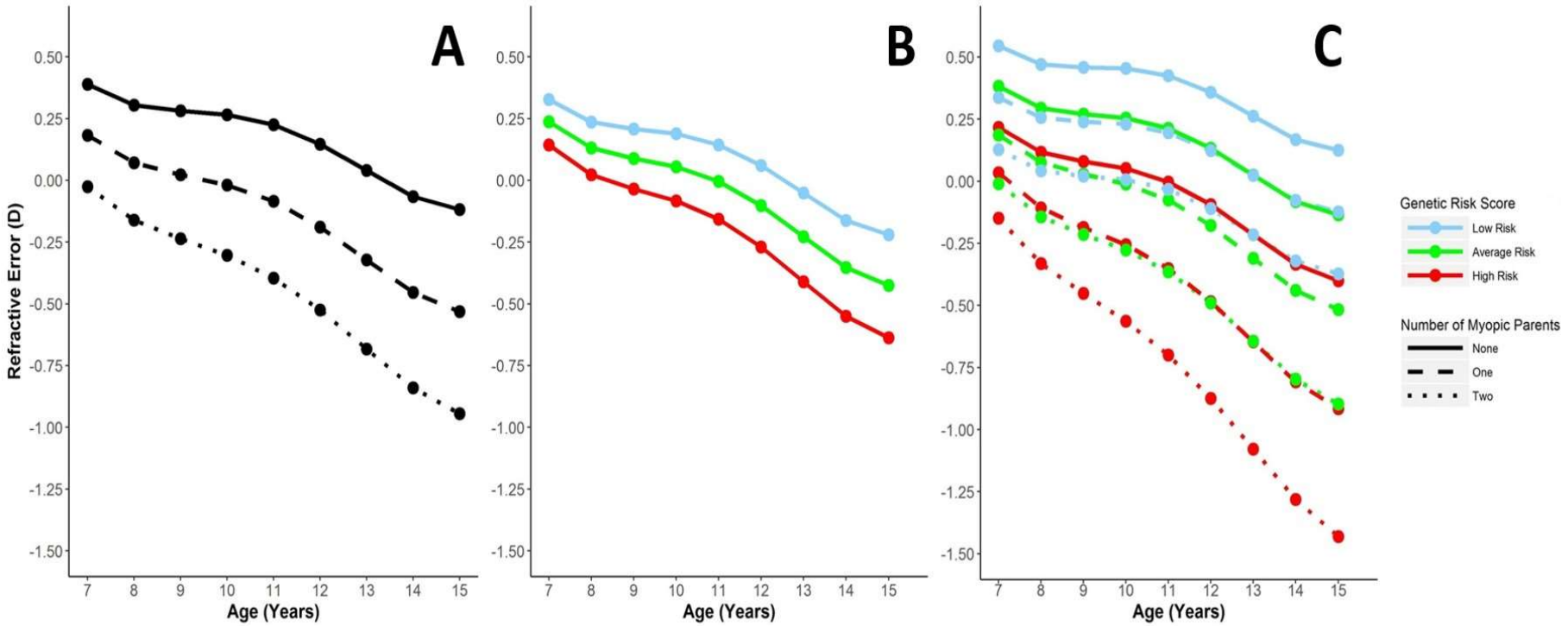


Figure 4.5. Refractive trajectories predicted using (A) number of myopic parents, (B) genetic risk Z score, and (C) a combined model with high, average, and low risk genetic risk categories for children with 0, 1, or 2 myopic parents. In (B) and (C), the high and low genetic risk categories correspond to children with a genetic risk score 1 standard deviation above or below the mean, respectively. This figure was created by using data from 2885, 2960, 2918, 2852, and 2368 children who attended at the ages of 7, 10, 11, 12, and 15, respectively (these children were all part of a subset of the full cohort, comprising of a total n = 3047 participants, who attended ≥3 research clinic visits). Adapted from Ghorbani Mojarrad et al. (2018).

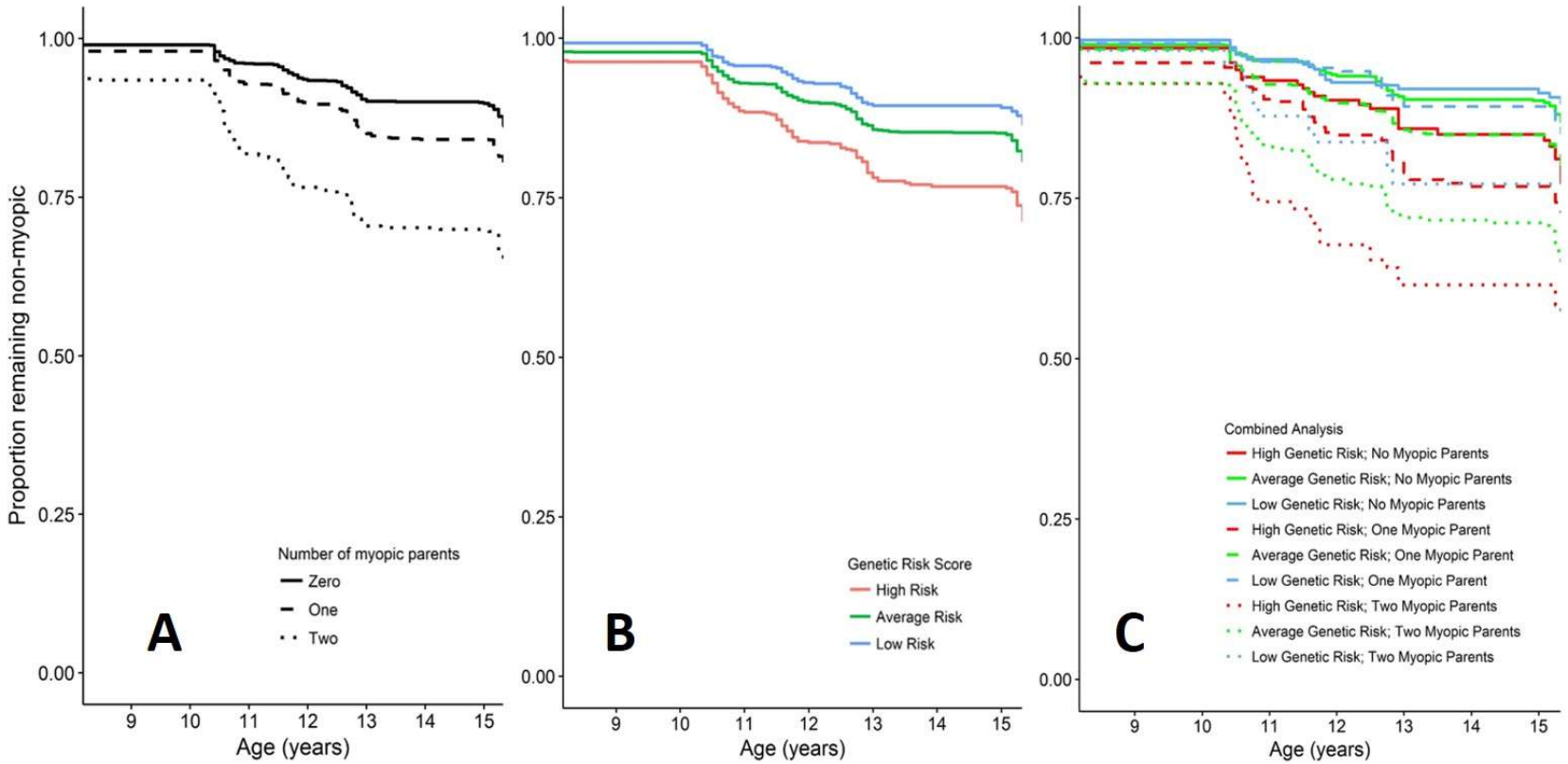


Figure 4.6. Survival curves for remaining non-myopic across the 9-15 year age range as a function of (A) number of myopic parents, (B) genetic risk score, and (C) a combined model with genetic risk score and number of myopic parents. In (B) and (C), the high and low genetic risk categories correspond to children with a genetic risk score 1 standard deviation above or below the mean, respectively. Adapted from Ghorbani Mojarrad et al. (2018).

4.4 Discussion

The refractive error of children in the ALSPAC cohort was assessed longitudinally between the ages of 7 and 15 years-old. These data were studied in order to test two closely-related hypotheses. The first was that using the genetic risk score would improve the prediction accuracy of refractive error in children beyond that of knowing their number of myopic parents. The second hypothesis was that predicting incident myopia based on knowing the number of myopic parents would also be enhanced by considering a genetic risk score. The results demonstrated that the genetic risk score did indeed improve prediction accuracy beyond that obtained through knowledge of the number of myopic parents (both at 7 and 15 years old). Thus, the first hypothesis was supported. Furthermore, in support of the second hypothesis, prediction of incident myopia also improved when both predictors were combined, compared to prediction based on knowing the number of myopic parents alone ($P = 5.1x10^{-10}$). This suggests that the number of myopic parents and the genetic risk score are at least partially independent of one another.

Despite the improved predictive performance when combining the genetic risk score and the number of myopic parents, the highest R^2 value achieved was ~7%. This is still too low to have clinical utility. This R^2 value demonstrates the difficulty in predicting the development of refractive error in an individual compared to explaining the refractive trajectories of a group of individuals. This is the same phenomenon which causes 95% prediction intervals to be much wider than 95% confidence intervals.

Comparison of the two different models for calculating genetic risk scores (the unweighted allele score model and the weighted allele score model) indicated that using the weighted allele score provided better predictive performance, although the improvement only had statistical support ($P < 0.05$) for the analysis at the age of 15. Although unweighted allele scores are simpler to determine, the weighted genetic risk score had the benefit of assigning more importance to variants with larger effect sizes.

Commonly-occurring genetic variants have been reported to explain approximately 39% of the variance in refractive error (known as “SNP heritability”), and therefore in theory, genetic prediction is capable of achieving this level of accuracy (Guggenheim et al. 2017; Shah et al. 2018). The reason for the poor performance of the genetic risk score in this

circumstance was likely due to: (1) the omission of risk variants which did not reach genome-wide significance in the CREAM and 23andMe, and (2) using imprecisely-estimated effect sizes for the 149 variants that were included. Both of these limitations could be improved by conducting a GWAS for refractive error in a larger sample, as well as including more variants in the analyses. This is supported by the results of Tedja et al., (2018), who reported a prediction accuracy of 7.8% for refractive error in an independent sample of adults when using over 7,000 genetic variants. Therefore, in future analyses to predict refractive error using genetic risk scores, a wider selection of genetic variants should be used to maximise the prediction accuracy. GWAS analyses on large available datasets would therefore be necessary to create genetic risk scores with a large number of variants included. This topic is the focus of Chapter 7, in which genetic risk scores are created using hundreds of thousands of genetic variants.

There are a further two additional reasons that could account for the relatively poor performance of the genetic risk score. Firstly, refractive error was assessed using non-cycloplegic auto-refraction, introducing a measurement error and thereby reducing the prediction accuracy. Secondly, refractive development continues into adulthood; therefore genetic risk scores created using GWAS summary statistics from adult populations are likely to perform better at predicting refractive error in adults than in children. Both the unweighted and weighted genetic risk score demonstrated improved accuracy at the age of 15, but not at age 7. This is likely to be due to phenotype immaturity at the age of 7. Extrapolating from these results suggests that refractive error prediction in this sample of ALSPAC participants will be more accurate when they have reached adulthood.

Figure 4.3 illustrates the difference in genetic risk score between individuals with different numbers of myopic parents. Although this graph suggests that myopic parents carry more genetic risk variants and pass these on to their children (which was supported by the statistical analyses), the extensive overlap across groups is also very apparent. Interestingly, the increase in the average genetic risk score with 0 vs.1, and 1 vs. 2 myopic parents of 0.15, and 0.36, respectively, suggests there is an approximately linear increase in genetic risk with the increase in the number of myopic parents. This would be consistent with a polygenic model of inheritance of refractive error.

It can be speculated that the reason for the improved predictive performance when combining the two predictors – number of myopic parents and genetic risk score – arises because number of myopic parents not only captures information about genetic risk, but also the risk from environmental factors. This suggests that myopic parents not only pass on their predisposing genes for myopia to their children, but also raise their children in a relatively myopia-inducing environment or manner. This suggestion has also been used to explain why estimates of the heritability of refractive error are higher in sibling-sibling comparisons vs. more distant relatives (Chen et al. 2007b). Overall, this indicates that using both a genetic risk score and knowledge of the number of myopic parents would be beneficial for myopia risk prediction. This would be possible at birth, as both of these predictive variables would not change throughout the child's lifetime.

There are several limitations for this work that should be taken into consideration. The ALSPAC children were all born within a few months of each other and recruited from the same geographic area. Results with a wider age range or from a larger geographic area may differ depending on the range and levels of myopia-inducing environment exposures (e.g. time outdoors and local schooling inconsistencies). Moreover, the analysis was restricted to participants of European ancestry; the results found would have likely have been worse for individuals with different ethnic backgrounds (Canela-Xandri et al. 2016). This topic is investigated further in Chapter 8.

Measuring refractive error with non-cycloplegic autorefraction typically leads to an age-specific bias in estimation of refractive error (Williams et al. 2008a). This measurement error would be expected to reduce the accuracy of genetic prediction of refractive error (Guggenheim et al. 2015), although without invalidating the comparison made between models using different predictors. Additionally, number of myopic parents was assessed with a simple questionnaire at a single time-point. Consequently, parental myopia may have been inferred incorrectly (with some parents misunderstanding the questions, or completing the questionnaire after undergoing refractive surgery) reducing the reliability of these results.

In conclusion, there was an improvement in the performance for predicting refractive error and incident myopia when information about genetic susceptibility and parental myopia was combined. Nevertheless, the predictive performance of even the best

model was poor ($R^2 \leq 7\%$). This suggests that further steps will be required in order to improve the accuracy of genetic risk scores, such as (1) including more genetic variants, and (2) estimating beta coefficients in a larger GWAS sample size.

5 Genome-Wide Association Study for Autorefraction-Measured Refractive Error in UK Biobank Participants

5.1 Introduction

In Chapter 4, the prediction of refractive error and myopia using genome wide significantly associated SNPs demonstrated lower accuracy to previously published data (7.8% accuracy from Tedja et al., (2018)), and performed less accurately in predicting myopia than using the number of myopic parents. This inferior prediction accuracy is likely due to the restricted number of genetic variants used to develop the genetic risk score (Dudbridge 2013), and therefore conducting a GWAS to obtain summary statistics for a greater number of genetic variants would be beneficial.

To carry out a GWAS for refractive error, a dataset of genotyped individuals with common ancestry who have had their refractive error measured is required. As discussed in Section 2.1, UK Biobank has released genetic data for a cohort of approximately 500,000 individuals from the UK, 23.3% of whom had refractive error measurements taken. Thus, these participants would be used for obtaining GWAS summary statistics that could be applied for genetic prediction.

As with any GWAS, it is important to evaluate the results obtained to assess if they are reliable, as GWAS results can give spurious associations for reasons such as unknown sample biases and uncontrolled confounders (Korte and Farlow 2013; Thomson and McWhirter 2017). The gold standard method of validating the reliability of GWAS results is through replication (Bush and Moore 2012), requiring the initial dataset to be divided into a test and replication set. However, this approach has the disadvantage of reducing the size of the sample used in the discovery of associated variants. Therefore, as an alternative, here I used the largest available UK Biobank sample for the GWAS for refractive error, and the results were compared to previously published associations. Note that a key disadvantage of this approach was that novel associations, by definition, could not be validated using existing published GWAS results.

At present, the two largest GWAS for refractive error or myopia were reported by Pickrell et al. (2016), and by the CREAM consortium and 23andMe (Tedja et al. 2018). The GWAS performed by Pickrell et al. (2016) was for the binary phenotype 'self-

reported near-sightedness' (Yes/No) from customers of 23andMe Inc. (Section 1.4.3). The GWAS reported by the CREAM consortium was a 'mega-analysis' of two datasets (a GWAS meta-analysis for age-of-onset of myopia carried out by 23andMe, and a GWAS meta-analysis for refractive error in 37 separate study samples carried out by CREAM). The respective sample sizes were 191,483 and 160,420 participants.

Here, a GWAS on the full dataset of 95,505 European participants from UK Biobank (who passed quality control filters and exclusion criteria) was performed, and the loci identified from this GWAS were compared to previously reported loci, to assess the concordance of the results. The hypothesis tested was that the GWAS performed on the UK Biobank participants would detect many loci associated with refractive error, and that these would be similar in their magnitude and direction of association with the phenotype to those of variants published previously.

5.2 Methods

5.2.1 Participant Selection

The UK Biobank dataset was used for this analysis. A total of 117,279 participants had refractive error information available (Cumberland et al. 2015). The refractive error phenotype ("Autorefraction MSE") was calculated as the mean spherical equivalent averaged between the two eyes, as described in Section 2.1.3.

To ensure participants were of a similar genetic background, the principal component analysis (PCA) results from Bycroft et al. (2018) were used (see Section 3.1.2). Participants were classified as having European ancestry if their first 20 principal components (PCs) were within ± 10 standard deviations of the mean for all those with self-declared "white British" ethnicity. Thus, individuals with ambiguous self-reported ethnicity (e.g. reported "prefer not to say" or "other") may still have been included in the analysis if they were clustered within the first 20 PC thresholds. After applying other participant quality control filters and the appropriate exclusion criteria (as described in Section 3.1.4), 95,505 participants were left who had sufficient data to conduct the GWAS for Autorefraction MSE. Note that as BOLT-LMM software was used (see below and Section 3.1.5), these participants were not filtered to remove related individuals, as BOLT uses a genetic relationship matrix to account for familial and cryptic relatedness.

5.2.2 GWAS for Autorefraction MSE

As stated above, BOLT-LMM software (Loh et al. 2015b) (version 2.3.2) was used to conduct the GWAS for Autorefraction MSE. Quality control filters for genetic variants used in this analysis are listed in Section 3.1.4. A QQ plot was created to allow comparison of the distribution of the observed association test statistics to that expected, with the corresponding lambda gc (λ_{gc}) value calculated to assess if there was any genomic inflation (Section 3.2.1). The LDscore regression intercept (Bulik-Sullivan et al. 2015b) was calculated to assess how much of this inflation was due to polygenicity (Section 3.2.2).

5.2.3 Comparison to GWAS Summary Statistics in Published Literature

The GWAS summary statistics for Autorefraction MSE were compared to those publicly available from the two studies mentioned in the introduction (Pickrell et al. 2016; Tedja et al. 2018). Comparison was performed regarding three different aspects of the summary statistics:

1. The location of genome wide significant associations, i.e. were the same lead variants (or variants at the same locus and in high LD with previously-associated variants) replicated in all three GWAS analyses?
2. Direction of effect. The direction of effect at each locus would be expected to be the same across all 3 GWAS analyses. For example, if the literature reports the myopia-predisposing risk allele of a specific variant as 'A', the risk allele from the GWAS for Autorefraction MSE would also be expected to be the 'A' allele for this variant. A variant with a discordant direction of effect would not constitute a valid replication even if the variant was found to be associated with refractive error in 2 or more of the 3 GWAS analyses.
3. The estimated contribution ("effect size") of each variant to the phenotype. Effect sizes were compared to see how well these correlated across different GWAS.

Only the top 50 associated variants were available from the GWAS reported by Pickrell et al. (2016) due to publication restrictions. This meant that a direct comparison of all of the genetic loci identified was not possible between the three studies.

Due to the different GWAS analyses sometimes highlighting different lead variants at the same locus, the LDlink online reference tool (https://ldlink.nci.nih.gov/?tab=home) was used to identify nearby variants in high LD with lead variants that could serve as a surrogates for one another. Variants within 1 million base pairs of one another, and demonstrating high LD ($r^2 \geq 0.9$), were accepted as belonging to the same locus. Miami plots for the different studies were created to compare the genetic loci from pairs of GWAS analyses.

For variants at the same locus, risk alleles from the GWAS for Autorefraction MSE were compared to those from the previous GWAS publications to ensure they matched. If a different variant was identified as a lead SNP at the same locus, the LDlink online reference tool (see above) was used to identify which alleles were most commonly inherited together. Should the two variants identified at the loci in the two studies be in high LD (as explained above), and demonstrate consistency in the alleles inherited together, then they were deemed to have the same risk allele attributed. In other words, if the risk allele for the SNP identified in one GWAS was commonly inherited alongside the other risk allele for the SNP identified in the other GWAS, they would be categorised as the same loci. For example, if allele 'SNP_X' from GWAS #1 had an LD r^2 of 0.9 with allele 'SNP_Y' from GWAS #2 located 50,000 base pairs away, they would be classed as the same locus. Should the two alleles have a lower r^2 value or be more than 1 million base pairs apart, or have the opposite allele stated as the risk allele, then they would be categorised as different loci.

Effect sizes were compared by calculating the correlation of effect sizes of replicated lead variants from the GWAS of Autorefraction MSE and the CREAM consortium GWAS (Tedja et al., 2018). As Z scores were reported for the CREAM consortium meta-analysis, effect sizes from the GWAS for Autorefraction MSE were transformed to Z scores to allow comparison.

5.3 Results

5.3.1 GWAS for Autorefraction MSE

There were 587 variants associated with Autorefraction MSE at the conventional genome-wide statistical significance threshold of $P < 5x10^{-8}$. These variants were

distributed across 150 distinct genetic loci. The results are presented as a Manhattan plot in Figure 5.1. The QQ plot for the association statistics is shown in Figure 5.2. The genomic inflation factor λ_{gc} had a value of 1.26, indicating deviation from the expected value of 1.00, suggesting the possibility of an upward bias. However, the LD score regression intercept had a value of 1.03, indicating that the majority of this deviation was due to polygenicity.

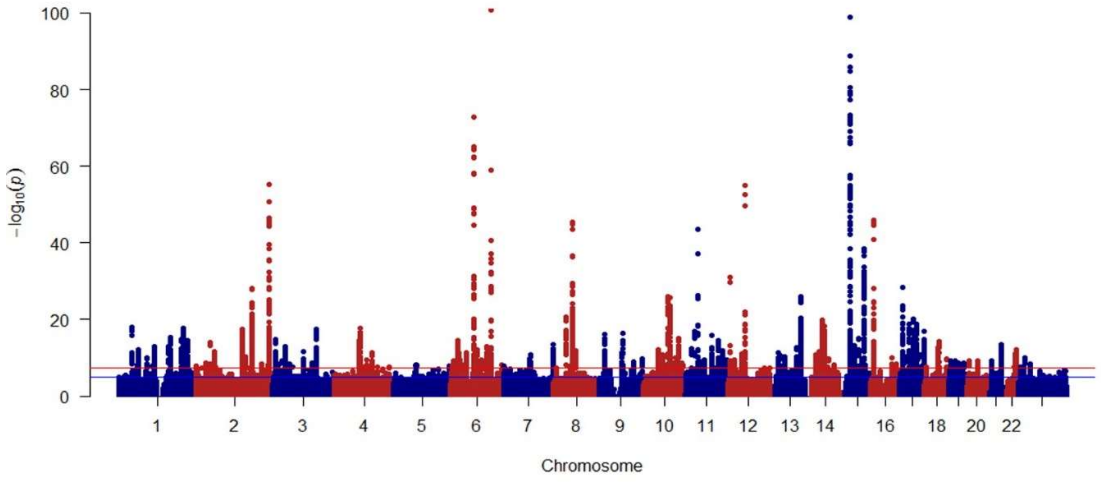


Figure 5.1 Manhattan plot demonstrating the results from a GWAS for Autorefraction MSE in 95,505 participants from UK Biobank. The red and blue lines indicate the conventional thresholds for declaring genome-wide statistical significance ($P < 5x10^{-8}$), and suggestive significance ($P < 5x10^{-5}$), respectively.

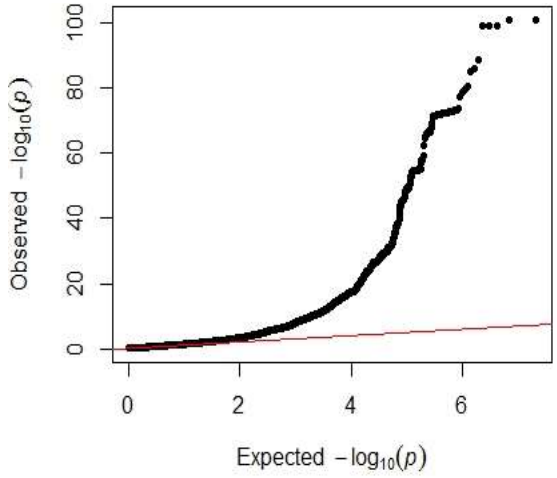


Figure 5.2 Quantile-quantile plot for the GWAS of Autorefraction MSE. λ_{gc} = 1.26.

5.3.2 Comparisons of GWAS results to other refractive error GWAS reports

Comparisons of the Manhattan plots of my GWAS for Autorefraction MSE and those of Tedja et al. and Pickrell et al. are shown as Miami plots in Figure 5.3 and Figure 5.4. In total, 100 of the 150 loci that reached genome wide significance ($5x10^{-8}$) in the Autorefraction MSE GWAS were replicated in the CREAM GWAS (Tedja et al., 2018). Full results are shown in Table 5.1. As described in the Methods, loci were deemed to replicate previously-reported associations if genome-wide significant variants in the two studies either (a) had the same rsID and location, or (b) were within 1cMB and in high LD ($r^2 \geq 0.9$). Table 1 also highlights those variants that were amongst the top 50 variants in the GWAS for self-reported myopia published by Pickrell et al. (2016). In total, all 50 of the 50 variants reported by Pickrell et al. replicated in the UK Biobank GWAS for Autorefraction MSE, with 49 of these same loci also being replicated in the CREAM consortium GWAS (Tedja et al., 2018), i.e. they were replicated in all three analyses. The single locus from Pickrell et al. that was not replicated in the CREAM analysis was on chromosome 11, replicating in the Autorefraction MSE GWAS. Table 5.1 also indicates which allele is the myopia-predisposing risk allele, and whether the direction of effect was consistent across GWAS analyses.

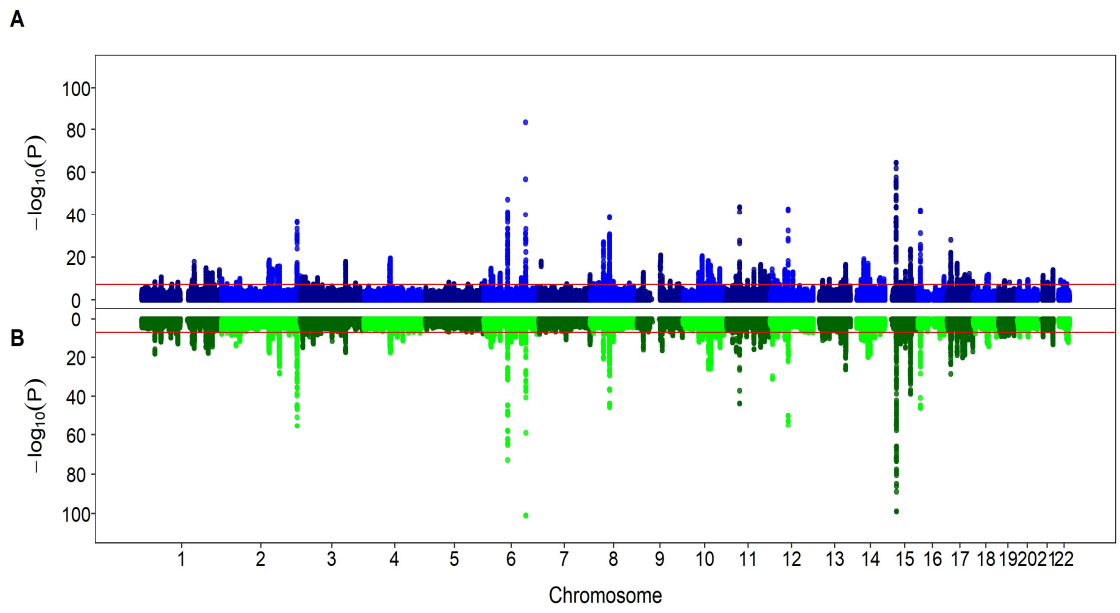


Figure 5.3 Miami plot comparing the results of the GWAS from Autorefraction MSE and CREAM consortium. The top panel shows data from the CREAM consortium analysis (adapted from Tedja et al. 2018), whereas the bottom panel shows data from the GWAS of Autorefraction MSE, an adaptation of Figure 5.1.

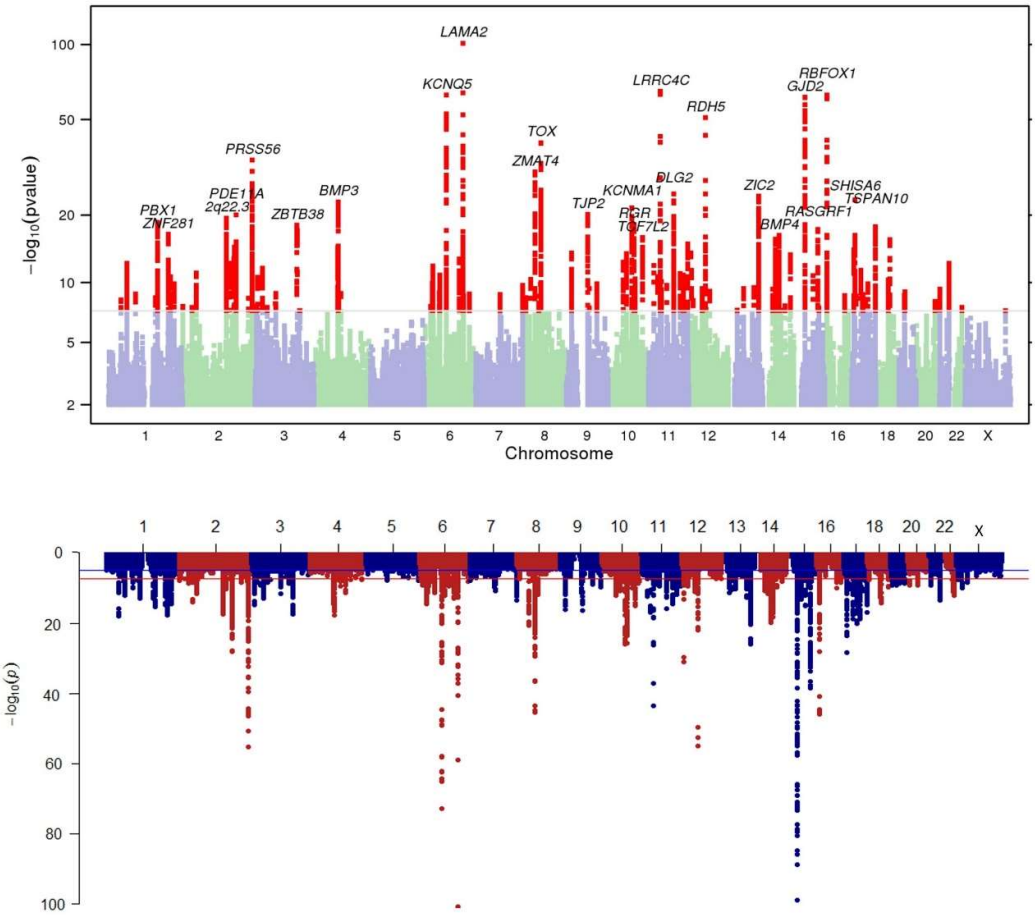


Figure 5.4 Miami plot comparing the results of the GWAS from Autorefraction MSE and self-reported myopia. The top panel shows data from Pickrell et al. (2016), and the bottom panel shows data from the GWAS of Autorefraction MSE, taken from Figure 5.1. The original data from the Pickrell et al. study was unavailable and therefore it should be noted that the alignment and scaling of this Miami plot are imprecise.

Table 5.1 (Overleaf) Variants exhibiting genome-wide significant association in the GWAS for Autorefraction MSE. Whether the variant replicated in the CREAM analysis (Tedja et al. 2018) and reported by Pickrell et al. (2016) as one of the top 50 loci is also indicated (8th & 9th column). The direction of the effect size and it's concordance between the three GWAS tests (if applicable) is shown in the 10th column.

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs2808511	1	200336096	**T**	**C**	-0.095	0.011	Y	Y	---
rs12028838	1	219778675	**T**	**G**	-0.092	0.012	Y		--
rs9787108	1	158033331	**C**	**T**	-0.089	0.012	Y	Y	---
rs112867366	1	113488343	**A**	**G**	-0.103	0.014	Y		--
rs579728	1	61185927	**A**	**G**	-0.083	0.012	Y	Y	---
rs12046000	1	91192396	**A**	**T**	-0.075	0.012	N		-
rs663431	1	108085902	**T**	**C**	-0.066	0.011	Y		--
rs1550094	2	233385396	**G**	**A**	-0.198	0.013	Y	Y	---
rs2695760	2	178835711	**A**	**G**	-0.130	0.012	Y	Y	---
rs62169542	2	146937226	**T**	**C**	-0.102	0.012	Y	Y	---
rs75120545	2	44271496	**T**	**C**	-0.278	0.036	Y	Y	---
rs41393947	2	56011517	**A**	**G**	-0.115	0.017	Y		--

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs576950451	2	157386310	**A**	**C**	-0.085	0.013	Y	Y	---
rs6736034	2	30468348	**C**	**G**	-0.076	0.013	N		-
rs13010104	2	208369213	**C**	**T**	-0.091	0.015	N		-
rs13382950	2	227956067	**C**	**T**	-0.067	0.011	Y		--
rs10188860	2	55237	**T**	**C**	-0.131	0.023	N		-
rs2110399	2	60516388	**A**	**G**	-0.066	0.012	N		-
rs62182439	2	172799794	**A**	**G**	-0.075	0.013	Y	Y	---
rs72772496	2	16430349	**C**	**T**	-0.120	0.022	Y		--
rs1582874	3	141115219	**T**	**C**	-0.100	0.012	Y	Y	---
rs502410	3	8182658	**C**	**A**	-0.094	0.012	Y		--
rs35135108	3	41166291	**C**	**T**	-0.088	0.012	N		-
rs9824877	3	98896242	**A**	**G**	-0.103	0.015	N		-

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs17497118	3	127334598	**G**	**A**	-0.087	0.015	N		-
rs2304577	3	53778621	**A**	**G**	-0.098	0.017	Y		--
rs1700943	3	24231030	**C**	**G**	-0.068	0.012	Y	Y	---
rs1568072	3	11041606	**G**	**A**	-0.079	0.014	N		-
rs4685282	3	15998835	**A**	**G**	-0.065	0.012	Y		--
rs35667547	3	64547477	**C**	**G**	-0.097	0.018	N		-
rs74764079	4	81952637	**A**	**T**	-0.311	0.036	Y	Y	---
rs536204902	4	120915393	**G**	**T**	-0.084	0.012	N		-
rs13107325	4	103188709	**C**	**T**	-0.140	0.022	N		-
rs59473955	4	89757082	**T**	**C**	-0.077	0.014	Y	Y	---
rs147792504	4	174336590	**A**	**T**	-0.231	0.042	Y	Y	---
rs554831360	4	138198041	**C**	**A**	-0.238	0.043	N		-

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs62395084	5	178338733	**G**	**A**	-0.095	0.018	Y		--
rs12193446	6	129820038	**A**	**G**	-0.422	0.020	Y	Y	---
rs7744813	6	73643289	**A**	**C**	-0.216	0.012	Y	Y	---
rs4145443	6	22068174	**T**	**G**	-0.092	0.012	Y		--
rs3812112	6	116444607	**A**	**T**	-0.088	0.012	Y	Y	---
rs12202798	6	84339255	**G**	**A**	-0.089	0.013	N		-
rs418092	6	28533946	**T**	**C**	-0.086	0.012	Y	Y	---
rs2746646	6	50818239	**T**	**C**	-0.081	0.011	Y	Y	---
rs6931604	6	98578215	**T**	**C**	-0.075	0.012	N		-
rs6917995	6	26327814	**C**	**T**	-0.073	0.011	N		-
rs2327222	6	10036083	**C**	**T**	-0.074	0.012	N		-

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs2326838	6	6901663	**A**	**G**	-0.070	0.012	N		-
rs11751433	6	163792854	**A**	**G**	-0.067	0.012	N		-
rs62485858	7	158893020	**T**	**G**	-0.120	0.014	Y		--
rs4278108	7	84317206	**G**	**A**	-0.093	0.014	Y		--
rs11764212	7	2067593	**C**	**A**	-0.066	0.012	N		-
rs72621438	8	60178580	**C**	**G**	-0.172	0.012	Y	Y	---
rs869422	8	40723970	**A**	**G**	-0.136	0.014	Y	Y	---
rs4738094	8	71423744	**A**	**G**	-0.086	0.012	Y		--
rs7465621	8	53333700	**G**	**A**	-0.103	0.018	N		-
rs10100265	8	10633159	**C**	**A**	-0.066	0.012	Y		--
rs4738828	8	61727587	**G**	**T**	-0.067	0.012	Y		--
rs7042950	9	77149837	**G**	**A**	-0.117	0.014	Y		--

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs1340044	9	18362105	**A**	**T**	-0.099	0.012	Y	Y	---
rs11145746	9	71834380	**A**	**G**	-0.099	0.014	Y	Y	---
rs72773790	9	129109080	**T**	**C**	-0.080	0.012	Y		--
rs10978697	9	109763587	**G**	**A**	-0.109	0.018	N		-
rs11596489	10	79054560	**T**	**C**	-0.128	0.012	Y	Y	---
rs4517452	10	86021024	**C**	**T**	-0.132	0.013	Y	Y	---
rs17747324	10	114752503	**C**	**T**	-0.111	0.014	Y	Y	---
rs4491171	10	49412862	**G**	**C**	-0.094	0.013	Y	Y	---
rs1658471	10	60291911	**A**	**T**	-0.078	0.012	Y	Y	---
rs80325284	10	102633779	**G**	**C**	-0.134	0.021	Y		--
rs36212732	10	124215198	**G**	**A**	-0.085	0.014	N		-
rs4747241	10	74036429	**T**	**C**	-0.070	0.012	Y		--

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs1254701	10	126810140	**A**	**G**	-0.090	0.015	N		-
rs6584962	10	111689904	**A**	**G**	-0.093	0.016	Y		--
rs541791855	10	94983203	**C**	**A**	-0.068	0.011	Y		--
rs999951	10	90036367	**C**	**G**	-0.076	0.014	Y		--
rs11602008	11	40149305	**T**	**C**	-0.218	0.016	Y	Y	---
rs7944541	11	30054610	**T**	**G**	-0.116	0.014	N		-
rs4943906	11	84638826	**T**	**A**	-0.097	0.012	Y	Y	---
rs10895869	11	105600358	**A**	**C**	-0.096	0.012	Y	Y	---
rs1550870	11	18751041	**C**	**T**	-0.085	0.012	N	Y	--
rs6421566	11	117671398	**A**	**G**	-0.080	0.011	Y	Y	---
rs2195526	11	119236751	**T**	**C**	-0.098	0.015	N		-
rs1015053	11	131933511	**A**	**T**	-0.074	0.012	Y	Y	---

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs198442	11	61506468	**T**	**C**	-0.073	0.012	N		-
rs227061	11	108205329	**A**	**G**	-0.072	0.011	N		-
rs2172998	11	43290063	**C**	**A**	-0.074	0.012	Y		--
rs654169	11	128691920	**A**	**G**	-0.070	0.013	Y		--
rs1938929	11	86334064	**C**	**T**	-0.069	0.013	N		-
rs3138142	12	56115585	**C**	**T**	-0.213	0.014	Y	Y	---
rs5442	12	6954864	**A**	**G**	-0.261	0.022	Y	Y	---
rs12146879	12	46408489	**G**	**A**	-0.092	0.013	Y		--
rs12423535	12	22533320	**C**	**A**	-0.077	0.012	Y		--
rs10842914	12	9275778	**T**	**C**	-0.078	0.013	Y	Y	---
rs2160729	12	14071844	**C**	**T**	-0.079	0.013	Y		--
rs892251	13	100647612	**G**	**A**	-0.126	0.012	Y	Y	---

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs837344	13	101186056	**T**	**C**	-0.098	0.012	Y		--
rs12853508	13	85607848	**T**	**G**	-0.097	0.013	Y		--
rs2281827	13	29001721	**C**	**T**	-0.094	0.014	Y		--
rs45502300	13	36246512	**G**	**A**	-0.243	0.037	N		-
rs7326825	13	50113450	**A**	**G**	-0.085	0.013	Y		--
rs1323971	13	94027893	**A**	**G**	-0.073	0.012	Y		--
rs2855530	14	54421917	**C**	**G**	-0.107	0.012	Y	Y	---
rs35320790	14	61108825	**C**	**A**	-0.103	0.011	Y		--
rs10483522	14	42275964	**T**	**C**	-0.100	0.015	Y	Y	---
rs2143975	14	33297398	**G**	**C**	-0.079	0.012	Y		--
rs74384554	14	74964903	**T**	**C**	-0.991	0.050	N		-
rs7149665	14	92590120	**C**	**T**	-0.092	0.016	Y	Y	---

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs3211166	14	69703158	**G**	**A**	-0.071	0.013	N		-
rs634990	15	35006073	**C**	**T**	-0.247	0.012	Y	Y	---
rs1961579	15	79380516	**A**	**G**	-0.154	0.012	Y	Y	---
rs7162310	15	63571234	**C**	**T**	-0.114	0.014	Y		--
rs75227249	15	48763008	**A**	**T**	-0.129	0.017	Y		--
rs893819	15	74229524	**G**	**A**	-0.077	0.012	N		-
rs62017256	15	50990965	**A**	**G**	-0.219	0.035	N		-
rs1112988	15	82318653	**A**	**G**	-0.075	0.013	Y		--
rs7188859	16	7460426	**C**	**T**	-0.175	0.012	Y	Y	---
rs28587148	16	67718563	**C**	**A**	-0.098	0.017	N		-
rs12919036	16	80423907	**G**	**A**	-0.069	0.012	Y		--
rs1868289	16	10215813	**T**	**G**	-0.069	0.013	N		-

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs2908972	17	11407259	**A**	**T**	-0.134	0.012	Y	Y	---
rs4794029	17	47280301	**C**	**T**	-0.117	0.013	Y		--
rs62067167	17	31251711	**T**	**C**	-0.139	0.015	Y	Y	---
rs1963456	17	54715143	**C**	**T**	-0.110	0.012	Y		--
rs9911460	17	79538841	**T**	**A**	-0.099	0.011	Y	Y	---
rs4793501	17	68718734	**T**	**C**	-0.079	0.012	Y		--
rs115152181	17	14136125	**A**	**T**	-0.078	0.012	Y		--
rs3785837	17	59468942	**A**	**G**	-0.088	0.014	N		-
rs876493	17	37824545	**G**	**A**	-0.075	0.012	Y		--
rs9038	17	75495397	**C**	**T**	-0.068	0.011	Y	Y	---
rs55754534	18	47433745	**C**	**G**	-0.129	0.016	Y	Y	---
rs7235709	18	42899939	**A**	**G**	-0.119	0.016	Y		--

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs3829640	18	72174980	**A**	**G**	-0.086	0.014	N		-
rs55728756	18	6460206	**T**	**C**	-0.067	0.012	N		-
rs55765017	19	19368264	**A**	**G**	-0.099	0.016	Y		--
rs8104875	19	8234677	**A**	**G**	-0.074	0.011	N		-
rs12462330	19	31806469	**T**	**C**	-0.073	0.012	N		-
rs77128495	19	48533700	**C**	**T**	-0.114	0.019	N		-
rs184784558	19	45434255	**A**	**T**	-0.101	0.018	N		-
rs6054512	20	6761512	**C**	**T**	-0.074	0.012	Y		--
rs4911405	20	32674967	**T**	**C**	-0.076	0.012	Y		--
rs8132840	21	47326747	**A**	**G**	-0.089	0.011	Y	Y	---
rs2229741	21	16340289	**T**	**C**	-0.074	0.011	Y		--
rs9330813	22	46364161	**A**	**G**	-0.090	0.013	N		-

Marker	CHR	POS	Effect Allele	Other Allele	BETA	SE	Replication in CREAM analysis?	Replication in Pickrell et al. analysis?	Direction
rs546593346	22	42189847	**G**	**T**	-0.086	0.013	Y		--
rs9623017	22	39959057	**A**	**G**	-0.081	0.015	N		-
rs17313971	23	20615249	**T**	**G**	-0.039	0.012	N		-
rs54266568	23	13978733	**I**	**D**	-0.084	0.019	Y		--
rs376420707	23	9292236	**A**	**T**	-0.095	0.012	N		-

The lead variants at all replicated loci showed a concordant direction of effect with previously-published GWAS analyses (Table 5.1).

Z scores for the 100 loci that replicated between UK Biobank and CREAM were compared (Figure 5.5). For this analysis all replicated loci were compared using the myopia-predisposing allele. The correlation was r = 0.745.

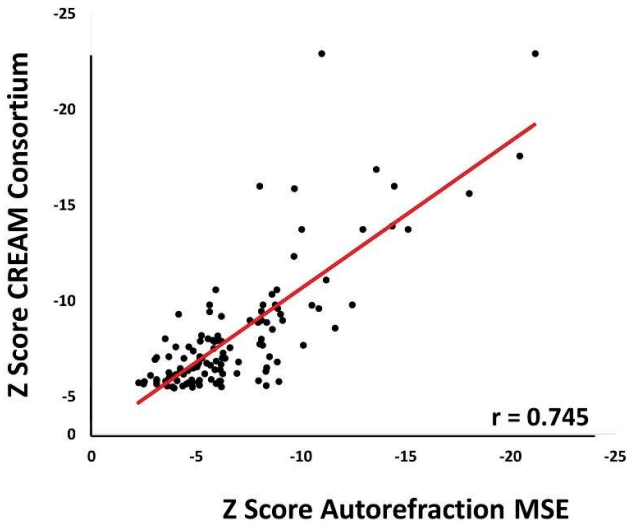


Figure 5.5 Scatter plot demonstrating the relationship between the direction and magnitude of association for lead variants at 100 loci displaying genome-wide significant association in a GWAS for Autorefraction MSE in UK Biobank and a GWAS for refractive error published by the CREAM consortium (Tedja et al., 2018). Effect size is quantified using the Z score.

5.4 Discussion

In this analysis a GWAS was performed for 95,505 individuals of European ancestry from UK Biobank who had their refractive error measured by autorefraction. This identified 150 loci, 100 of which directly replicated or were within 1MB of a variant in high LD in the CREAM consortium GWAS reported by Tedja et al (2018). Moreover, of the 50 top variants associated with self-reported myopia by Pickrell et al. (2016), all were replicated in the UK Biobank GWAS for Autorefraction MSE. A total of 49 of the 50 Pickrell et al. loci replicated in the CREAM GWAS analyses.

The GWAS for Autorefraction MSE summary statistics had a high λ_{gc} of 1.26, indicating possible inflation. However, the λ_{gc} value from the GWAS for Autorefraction MSE was not dissimilar to those reported previously: λ_{gc} = 1.13 in the CREAM analysis (Tedja et al. 2018), and λ_{gc} = 1.23 in the self-reported myopia analysis (Pickrell et al. 2016). There are

three reasons why a λ_{gc} value may be inflated i.e. greater than one: population stratification, relatedness (either through direct ancestry or cryptic relatedness), or polygenicity (Bulik-Sullivan et al. 2015b). The selection of individuals within the ±10 standard deviations of the first 20 PCs for European ancestry means that hidden population stratification is unlikely to be the cause. Moreover, related individuals were used in the GWAS for Autorefraction MSE, but the use of a genetic relatedness matrix created within BOLT software accounted for this. As the LDscore regression intercept reported a value of 1.03 (95% CIs 1.01-1.05), this indicated that the inflation was largely due to the polygenicity. Thus, this means that the UK Biobank summary statistics are unlikely to contain false positive findings caused by bias. Inflation of λ_{gc} values due to polygenicity is commonly seen with GWAS of large sample size.

As this experiment used all individuals available from UK Biobank in the GWAS for refractive error, with a view to maximising the sample size, there was no possibility of repeating the GWAS in other UK Biobank participants (i.e. there was no separate UK Biobank replication sample to validate the findings). However, replication in at least one other published GWAS for refractive error was considered as evidence of independent replication. 49 of the genome wide significant SNPs from the Autorefraction MSE GWAS were not identified in the publicly available results from the CREAM consortium (Tedja et al., 2018). If a random non-overlapping sample of UK Biobank participants had been kept aside for replication, it may have been possible to reproduce the association for the 49 SNPs not seen in the CREAM analysis. Furthermore, the loci from this GWAS that did not reach genome wide significance in the CREAM summary statistics data may have demonstrated association at the lower suggestive threshold (i.e. $5x10^{-5}$). It is likely that this lack of replication is due to the reduced power of the Autorefraction MSE GWAS to detect signals and associations for refractive error in comparison to the GWAS from Tedja et al., due to having a smaller sample size. Therefore, identifying a greater number of signals, including those identified by the CREAM consortium may be possible if the effective sample size can be increased.

The variant for self-reported myopia in the Pickrell et al. GWAS that did not replicate in the CREAM consortium results (but did show an association in the GWAS for Autorefraction MSE) was on chromosome 11; lead variant rs1550870. This variant is in a coding region for the Protein Tyrosine Phosphatase Non-Receptor Type 5 (*PTPN5*)

gene, also commonly known as the human *STEP* locus (an acronym for STriatal-Enriched protein tyrosine Phosphatase). This gene has been shown to be expressed in the brain, largely in the cerebral cortex (Lombroso et al. 1991). It has been shown to have an association with Alzheimer's disease, schizophrenia, and Parkinson's disease (Zhang et al. 2010; Carty et al. 2012; Kurup et al. 2015), with the respective diseases showing increased activity in this region alongside stress disorders (Yang et al. 2012). It is unknown how this variant may biologically impact pathways that induce myopia. In the GWAS for Autorefraction MSE, this variant had a P value of $4.5x10^{-13}$, similar to that in the report by Pickrell et al. (2016); P = $9.9x10^{-13}$. The locus was not genome-wide significant in the CREAM GWAS, which may be due to post-analysis quality control, or simply lack of significance after meta-analysis.

The most strongly associated variants showed evidence of good reproducibility (Figures 4.3-4.5). Not only did the loci mostly replicate between all three GWAS, but the top 9 loci in the UK Biobank analysis also appeared in the list of the top 9 loci in the other two GWAS analyses, (although not in the same descending order). The top variant in all three analyses was rs12193446 on chromosome 6. This variant lies in an intron of the *LAMA2* gene, and has previously been reported to have a strong association with refractive error and myopia (Verhoeven et al. 2013; Verhoeven et al. 2014; Li et al. 2015). It had the largest effect size for a genetic variant at -0.42D in the UK Biobank sample. This variant had the smallest P values of $1.1x10^{-118}$, $5.4x10^{-102}$, and $1.6x10^{-101}$ in the CREAM analysis, Pickrell et al. (2016) analysis, and Autorefraction MSE analysis respectively, alongside Z scores of -22.9, -21.7, and -21.9, respectively.

The correlation between the effect sizes for the top 100 variants that replicated can be seen more generally in Figure 5.5. All variants demonstrated the same direction of effect between UK Biobank and CREAM. Moreover, the correlation coefficient of 0.745 indicates that there was a generally strong positive correlation between the effect sizes estimated from each locus in the CREAM consortium GWAS (Tedja et al., 2018) and Autorefraction MSE GWAS. Although this value only relates to the top 100 loci in common between these two studies, all of the variants published by Pickrell et al. (2016) also had the same directionality for the identified risk alleles. That these 3 GWAS analyses identified a majority of the same loci, same risk alleles, and had similar effect

sizes increases the confidence that results of the GWAS for Autorefraction MSE are reliable and similar to previously published refractive error GWAS analyses.

It should be noted that the CREAM and Pickrell et al. GWAS analyses may have had partially overlapping samples (Pickrell et al. 2016; Tedja et al. 2018). All participants in the Pickrell et al. GWAS analysis were customers of 23andMe, as were 104,293 of the CREAM GWAS study participants. However, since the phenotypes studied in the Pickrell et al. and CREAM GWAS analyses were based on different questionnaire responses completed by 23andMe customers, the exact degree of overlap is unknown. Furthermore, the lack of availability of the full list of genome-wide significant loci in the Pickrell et al. GWAS limited the ability to compare associations across the 3 sets of GWAS summary statistics.

An additional limitation of the GWAS in UK Biobank participants was that only European participants were studied, meaning that the applicability to other ethnicities may be limited. It should also be noted that distinct phenotypes were analysed in the 3 GWAS studies (refractive error; myopia case-control status; age-at-onset of myopia). Although an adequate comparison was possible after transformation of effect sizes to Z scores, this difference in phenotype may have led to discrepancies in the results, e.g. leading to a reduced correlation coefficient.

In conclusion, a GWAS for Autorefraction MSE in UK Biobank participants identified 150 genome-wide significant loci. Amongst the top 50 GWAS variants identified previously by Pickrell et al. (2016), all were replicated in the Autorefraction MSE GWAS. Moreover, 100 of the loci were also replicated in the GWAS performed by Tedja et al. (2018), showing the same direction of effect and comparable effect sizes. Nevertheless, 49 genome wide significant SNPs did not show replication in previously published data at genome wide significance, likely due to an underpowered analysis of limited sample size. Therefore, identifying a method in which to increase the effective sample size used in GWAS analyses which could be used in downstream analysis to develop genetic risk scores may be beneficial.

6 Creation of Predictive Phenotypes and Comparison to Autorefraction MSE

6.1 Introduction

As discussed in Section 1.3.10 and Chapter 4, a limitation for the identification of genetic variants associated with refractive error and myopia is insufficient statistical power due to a limited sample size, which in turn can limit the accuracy of a genetic risk score (Dudbridge 2013). Because of this a GWAS for autorefraction MSE was performed; 95,505 UK Biobank participants of European ancestry with refractive error data were available, after applying quality control filters. These participants were studied in Chapter 5, in which a GWAS for refractive error (Autorefraction MSE) was carried out. However, as discussed in the literature review (Section 1.4.3), the largest refractive error GWAS meta-analysis conducted to date included 160,420 individuals (Tedja et al., 2018), more than in the Autorefraction MSE GWAS. A polygenic risk score derived using the top 7,307 variants identified in the Tedja et al. study had a prediction accuracy of R^2=7.8%, which is greater than the accuracy found in Chapter 4. The Tedja et al. study was notable because the authors meta-analysed summary statistics from a GWAS for refractive error and a GWAS for self-reported age of myopia onset. The resulting genetic risk score was more accurate (R^2=7.8%) than that previously reported in 2013 from a GWAS for refractive error in only 45,758 participants (R^2= 3.4%) (Verhoeven et al. 2013), with sample size likely to be the key factor (Dudbridge 2013). Thus, performing a meta-analysis of GWAS summary statistics for refractive error and a similar correlated trait may help improve the genetic risk score accuracy.

Myopia onset usually occurs between the ages of approximately 8 to 12 years, with some individuals developing myopia – to a much lesser extent – as young adults (Saw et al. 2002; Wojciechowski 2011; Pärssinen et al. 2014). In contrast, hyperopia is often established at birth or in early infancy, with spectacles often being prescribed before the age of 6 years-old (Mutti et al. 2007). Thus, refractive error is likely associated with the age of onset of spectacle wear (AOSW).

Moreover, as described in Section 1.2.3.2 (Vitale et al. 2009; Wen et al. 2013; Williams et al. 2015a), there is much evidence indicating an increase in the prevalence of myopia

over the past few decades, in Europe, the United States and East Asia. Therefore, adults born in later birth cohorts were at a higher risk of developing myopia and a more myopic/negative refractive error, on average. Thus, an individual's year-of-birth (YOB) is also associated with refractive error, and therefore likely explains some of the inter-individual variation of refractive error in the general population.

In this Chapter, these two variables, AOSW and YOB, were used to develop a statistical model for estimating refractive error. The model was optimised in a sample of UK Biobank participants who underwent autorefraction and who reported their AOSW. Once the model was optimised, it was employed to estimate the refractive error of UK Biobank participants who did not undergo autorefraction. A GWAS for predicted refractive error ('AOSW-inferred MSE'), using AOSW and YOB was performed. Tests were then carried out to determine if AOSW-inferred MSE was an effective surrogate phenotype for refractive error measured by autorefraction.

6.2 Methods

All statistical analyses were performed using R. All GWAS analyses were run using BOLT. A flow diagram of participant inclusion criteria for all samples used is shown in Figure 2.2.

6.2.1 Creation of a Predictive Model for Refractive Error Using Age of Spectacle Wear (AOSW) and Age

UK Biobank participants were stratified into two groups, dependent upon autorefraction-measured refractive error data availability. The sample of participants with refractive error data available was used to create and refine a model defining the relationship between AOSW and autorefraction-measured MSE. The model was then used to predict refractive error ('AOSW-inferred MSE') in participants whose refractive error was not known. The statistical model employed was a multivariable linear model that incorporated polynomial terms to account for non-linear relationships between refractive error and AOSW and between refractive error and YOB.

386,318 UK Biobank participants of European ancestry who self-reported their AOSW were used in this analysis (after applying quality control filters and exclusion criteria; Section 3.1.4). Of this sample, 98,870 participants ('Group 1') had a known refractive error from autorefraction, while 287,448 participants ('Group 2') did not have data

available for autorefraction. There was an approximately 1:3 ratio of participants with known average mean spherical equivalent data (Autorefraction MSE sample) and unknown refractive data (AOSW-inferred MSE sample).

AOSW was coded as a continuous variable. Self-reported Age was used as a surrogate for YOB (since for the UK Biobank sample, Age and YOB are very highly correlated). Age was also coded as a continuous variable in the model. A multivariable linear regression model was generated, which included these two predictor variables (AOSW and Age), along with Sex (coded as a binary variable). The distributions of AOSW and Age are shown in Figure 6.1. The 98,870 participants of European ancestry with known refraction were used to generate the model (see Figure 2.2 for more detail as to how these individuals were identified and extracted).

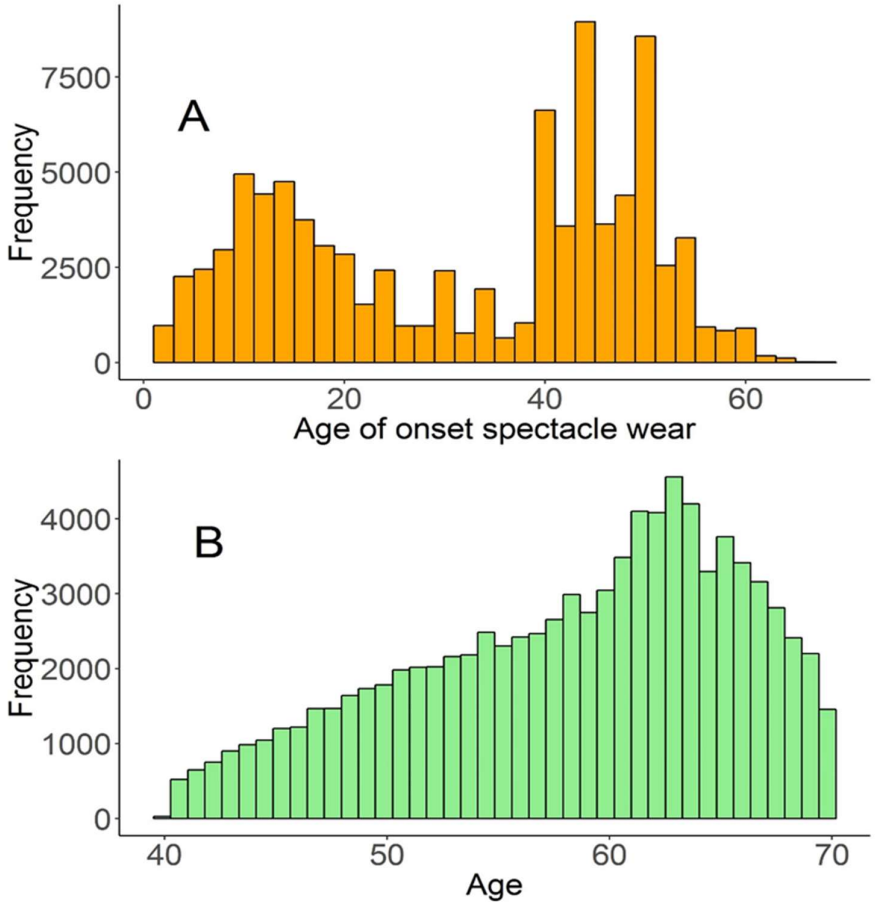


Figure 6.1. Distributions of AOSW (Panel A) and Age (Panel B).

An R script was written that contained a set of nested loops, in order to determine the optimal polynomial order for the variables Age and AOSW. The R function 'poly' was

used to adjust polynomial order. A mixture of lower-order and higher-order polynomial terms enabled the complex, non-linear relationship between refractive error and AOSW to be modelled accurately. An advantage of the R 'poly' function is that it generates orthogonal polynomials, which have the useful property of being uncorrelated; therefore avoiding unstable model fits as a result of multi-collinearity.

The linear regression model took the form:

$$avMSE \quad \sim \quad \alpha \; + (\beta_{1j} \times Age^{j}) \; + \; (\beta_{2k} \times AOSW^{k}) \; + \; (\beta_{3} \times Sex)$$

Equation 6.1. Estimation of refractive error. Here, j *and* k *represent polynomial orders.*

The optimal polynomial order was determined by empirical testing. Specifically, a likelihood ratio test was used to compare between a model with vs. without a one-step increase in polynomial order (e.g. Age^{j+1} compared to Age^{j}), for both Age and AOSW separately to determine optimal values for values *j* and *k*. If the likelihood ratio test between the previous polynomial order and new increased order yielded a significant improvement ($P < 0.05$) in model fit for the more complex model, the higher order model was selected. This approach was continued until increasing the polynomial order further did not yield a significant improvement in model fit. A full example of the code used to calculate polynomial orders is available in the appendix.

The model optimisation was done in a sample of participants with known refractive error. First, the Autorefraction MSE sample was split into 2 subgroups: a 'training' dataset and a 'test' dataset. These two groups were equal in size (containing 49,435 participants each). Regression models for selecting the optimal polynomial orders were fitted using the 'training' dataset. The optimal polynomial orders and regression coefficients from the training dataset model were then used to infer (i.e. predict) refractive error in the participants from the 'test' dataset based on their AOSW, Age and Sex. The coefficient of determination was calculated for the autorefraction-measured MSE vs. AOSW-inferred MSE relationship in the 'test' dataset.

6.2.2 Using the Optimised Model to Estimate Refractive Error in Participants With Unknown MSE

The R function 'predict' was used to apply the optimised model in the 287,448 participants in the 'AOSW-inferred MSE' sample. This step applied the optimised model

parameters to the AOSW, Age, and Sex of each participant in order to estimate their refractive error, i.e. to generate a value for the AOSW-inferred MSE phenotype of each individual.

6.2.2.1 Transformation of AOSW inferred MSE to a Normal Distribution ('AOSW norm MSE')

Once the AOSW inferred MSE phenotype had been calculated for each participant, consideration was also given as to whether transforming the distribution of the AOSW-inferred MSE phenotype would be beneficial to prediction accuracy. AOSW-inferred MSE values were transformed to a normal distribution using an inverse rank-based normalisation method (also known as van der Waerden transformation). Estimates of AOSW-inferred MSE values were listed in numerical order, and individuals were assigned a rank based on their position (thus each person was given a rank from 1 to 287,448 depending on their relative AOSW-inferred MSE value). Once ranked, 287,448 values from a simulated normal distribution (with a mean of 0 and a standard deviation of 1) were drawn, ranked in order as per the original AOSW-inferred MSE values, and then assigned to the dataset. This phenotype will be referred to as 'AOSW norm MSE'.

6.2.3 GWAS for 'AOSW-inferred MSE' and GWAS for 'AOSW norm MSE'

Once AOSW-inferred MSE and AOSW norm MSE had been calculated for the 287,488 individuals in the AOSW-inferred MSE sample (i.e. those participants that did not have refractive error data from autorefraction), a GWAS for each trait was carried out. This was done using BOLT-LMM software (Section 3.1.5). Quality control filters and exclusions were applied as described in Section 3.1.4. λ_{gc} was also calculated to test for genomic inflation, and the LDSC regression intercept calculated if λ_{gc} demonstrated any potential bias.

6.2.4 Comparison of GWAS Summary Statistics for 'Autorefraction MSE' vs. 'AOSW-inferred MSE' and 'AOSW norm MSE'

The GWAS summary statistics for Autorefraction MSE (i.e. the results from Chapter 5) and AOSW-inferred MSE or AOSW norm MSE were compared using two measures. Firstly, the genetic correlation between the two traits was calculated using LD score

regression (Section 3.2.3). Secondly, Pearson's correlation was calculated for the genetic 'effect size estimate' (i.e. GWAS regression coefficient) for markers strongly associated with one or both traits. For the latter analysis, markers were selected based on their P-value for association with the trait of interest.

6.3 Results

6.3.1 Determination of AOSW-inferred MSE model

Increasing the polynomial order for the variables AOSW and Age improved the fit of the regression model shown in Equation 6.1 for participants in the 'training' dataset. The R^2 of the models was calculated for each polynomial step until the optimal polynomial order was reached (Table 6.1 and 6.2 and Figure 6.2 and 6.3).

Polynomial Order for AOSW	R^2 of Model
1	0.167818
2	0.168501
3	0.209257
4	0.245416
5	0.258757
6	0.26354
7	0.264839
8	0.265659
9	0.269904
10	0.271878
11	0.272424
12	0.272684
13	0.272713
14	0.272726
15	0.272726

Table 6.1. The R^2 of a model for autorefraction-measured refractive error estimated from Age of Onset Spectacle Wear (AOSW) of different polynomial orders. There was no significant change in the R^2 after a polynomial order of 13 (i.e. likelihood ratio test P > 0.05).

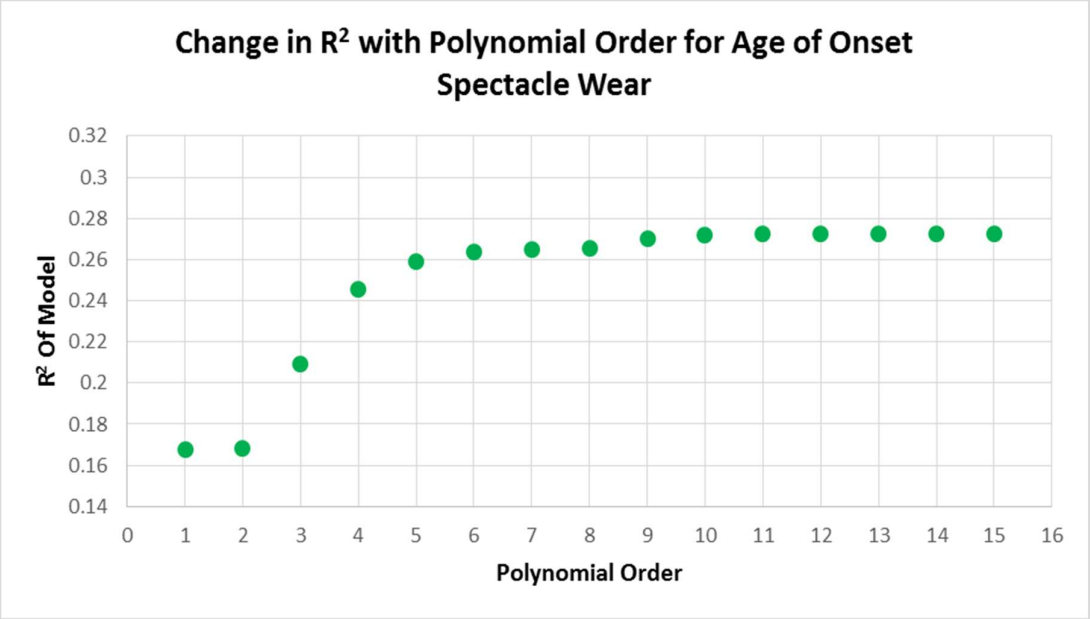


Figure 6.2. Graphical presentation of the changes in R^2 with different polynomial orders for Age of Onset Spectacle Wear (AOSW).

Polynomial Order	R^2 of Model
1	0.046504
2	0.046641
3	0.046762
4	0.047046
5	0.047093
6	0.047192
7	0.047182
8	0.047171
9	0.047175
10	0.047166

Table 6.2. The R^2 of a model for autorefraction-measured refractive error estimated from Age at different polynomial orders. There was no significant change in the R^2 after a polynomial order of 6 (i.e. likelihood ratio test $P > 0.05$).

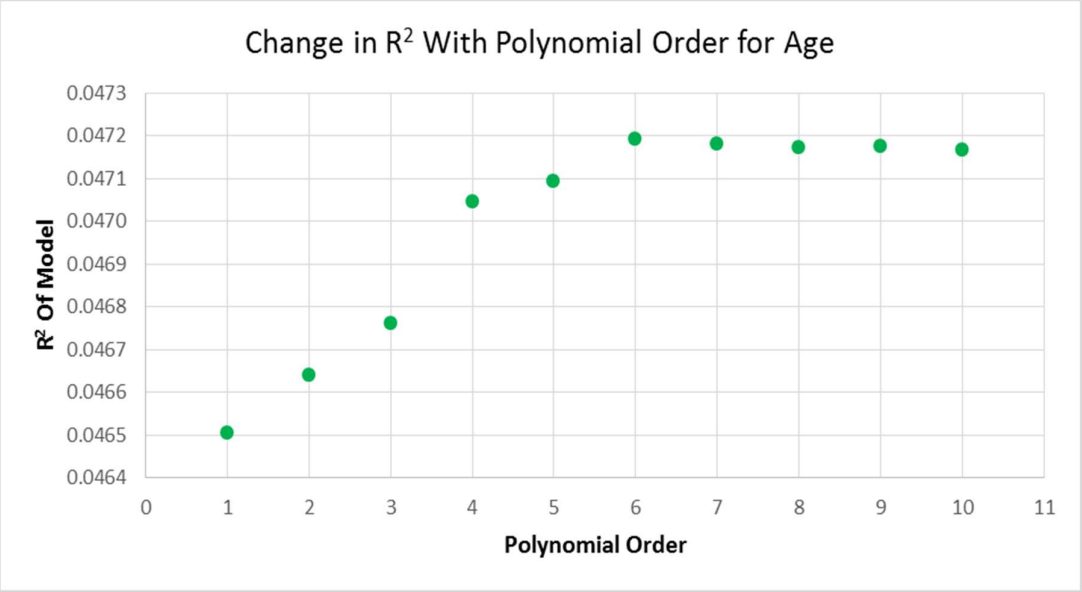


Figure 6.3. Graphical presentation of the changes in R^2 with different polynomial orders for Age.

The optimal polynomial orders were 13 and 6 for AOSW and Age, respectively. Consistently, AOSW achieved higher R^2 values than Age.

Using AOSW, Age, and Sex combined in a single predictive model with the selected polynomial orders, gave an R^2 of 0.30 (95% CI: 0.29 –0.31) in the independent 'test' dataset. The mean absolute error (MAE) for AOSW-inferred MSE was 1.54D (95% CI: 1.53D – 1.55D).

The mean value for AOSW-inferred MSE in participants with refractive data in the 'test' dataset was -0.30D, which was the same mean value for Autorefraction MSE. However, there was a difference between the standard deviations of AOSW-inferred MSE and Autorefraction MSE (1.55D and 2.80D, respectively) indicating that the distributions of these traits were different.

Figure 6.4 and Figure 6.5 demonstrate the distributions of refractive error for Autorefraction MSE and AOSW-inferred MSE.

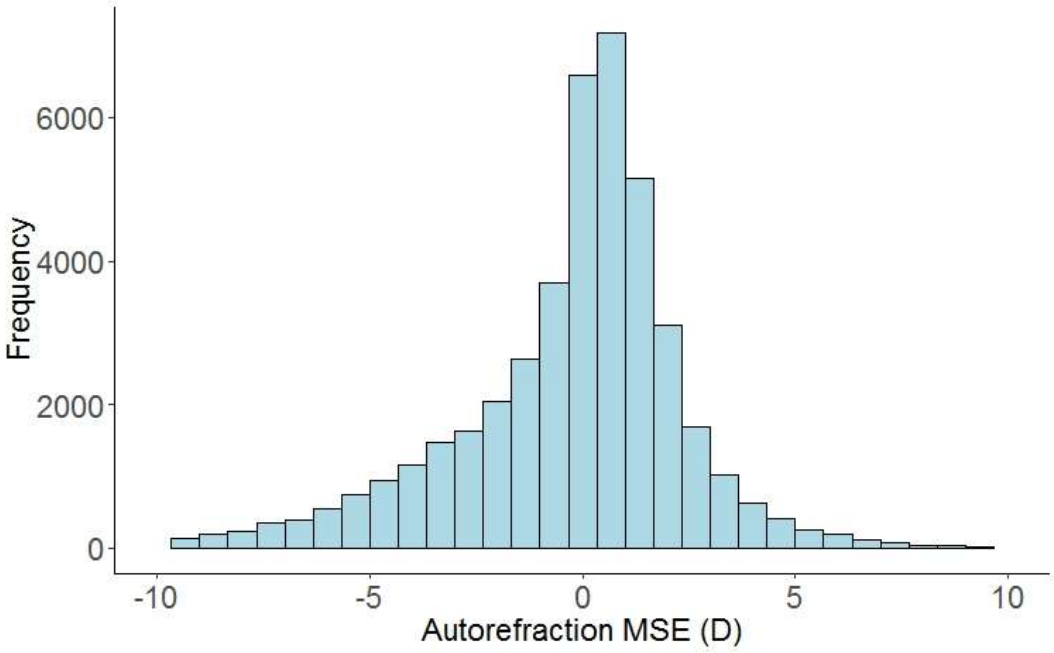


Figure 6.4. A histogram of Autorefraction-measured MSE in the 'test' dataset (N = 49,435).

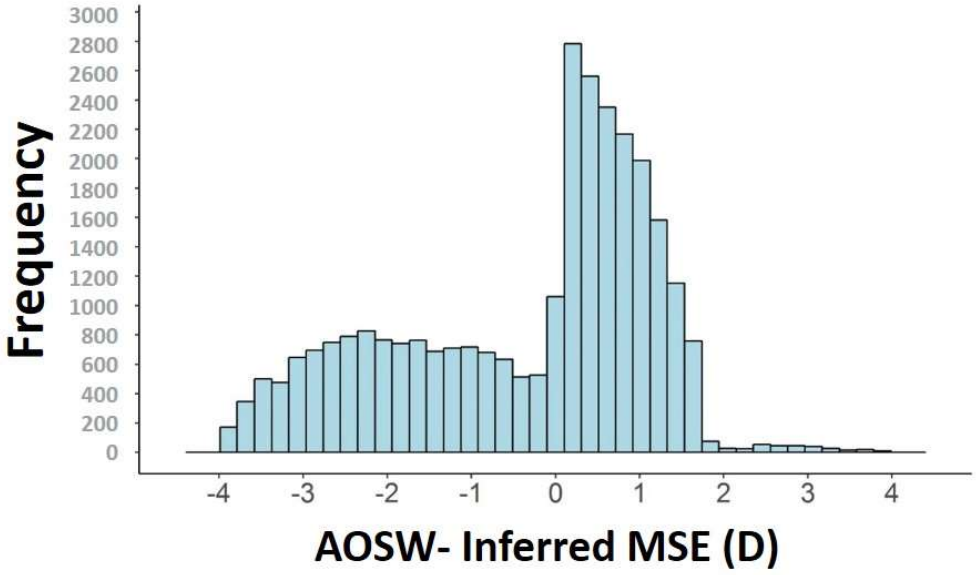


Figure 6.5. A histogram of AOSW-inferred MSE in the 'test' dataset (N = 49,435).

6.3.2 Transformation to a Normal Distribution: The 'AOSW norm MSE' trait

An inverse rank-based normalisation method was applied to transform AOSW-inferred MSE values to have a normal distribution with a mean of 0 and standard deviation of 1. The resulting trait was termed 'AOSW norm MSE'. The distribution of AOSW norm MSE is presented in Figure 6.6.

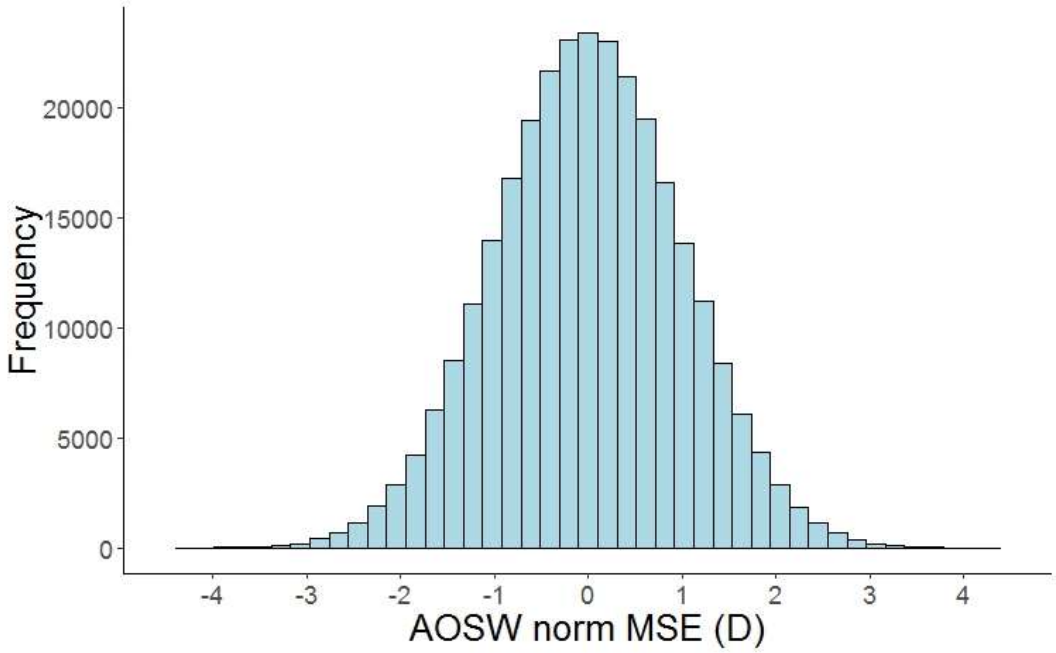


Figure 6.6. Histogram of AOSW norm MSE in individuals without refractive error data (N = 287,448).

For participants with Autorefraction MSE data available (N = 95,505), the AOSW norm MSE variable had a higher coefficient of determination than that of AOSW-inferred MSE (R^2 = 0.45 vs. R^2 = 0.30) in relation to the Autorefraction MSE. Thus, GWAS analyses were performed for both AOSW-inferred MSE and AOSW norm MSE.

6.3.3 Comparison of Results from GWAS for Autorefraction MSE, AOSW-inferred MSE, and AOSW norm MSE

A GWAS Manhattan plot for Autorefraction MSE has already been presented in Figure 5.1. It has been adjusted and fitted into the Miami plots in Figure 6.7 and Figure 6.8 to demonstrate the differences between the associations identified in the Autorefraction MSE GWAS compared to AOSW-inferred MSE and AOSW norm MSE, respectively.

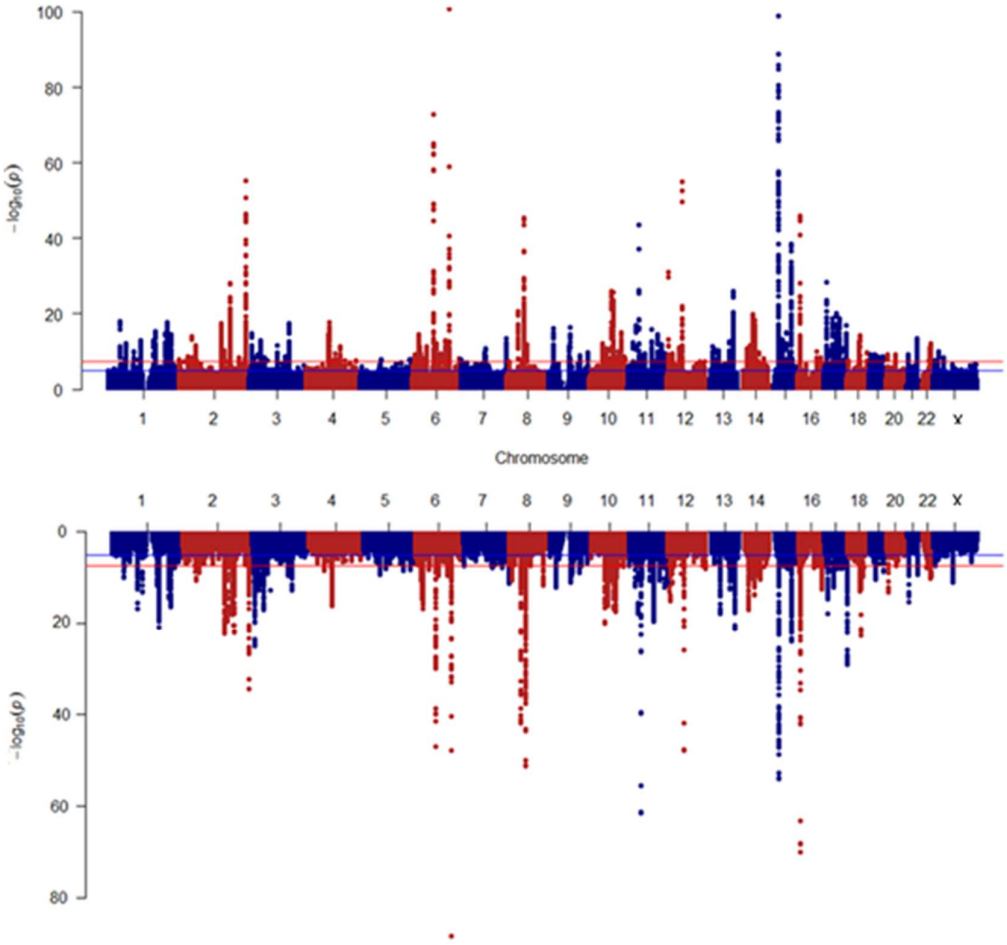


Figure 6.7 Miami plot comparing the GWAS results for Autorefraction MSE (top) and AOSW-inferred MSE (bottom). The blue and red lines indicate levels of suggestive significance and genome wide significance ($5x10^{-5}$ and $5x10^{-8}$), respectively.

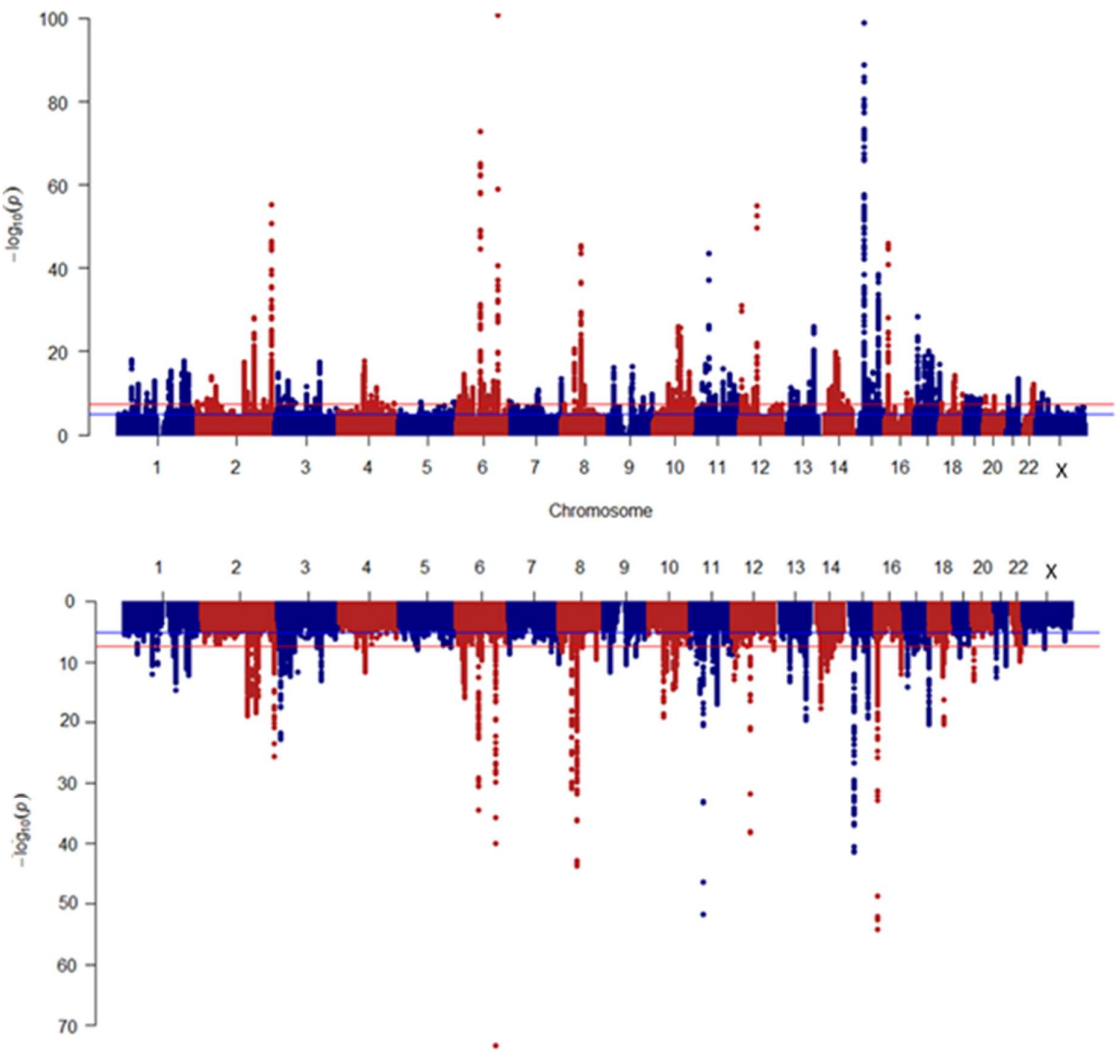


Figure 6.8 Miami plot comparing the GWAS results for True MSE (top) and Predicted Normalised MSE (bottom). The blue and red lines indicate levels of suggestive significance and statistical significance ($5x10^{-5}$ and $5x10^{-8}$), respectively.

6.3.4 Genetic Correlations

Table 6.3 demonstrates all the pairwise genetic correlations between the three traits.

All correlations were above +0.92.

	Autorefraction MSE	*AOSW-inferred MSE*	*AOSW norm MSE*
Autorefraction MSE	*1.00*	*0.92 (0.92 to 0.92)*	*0.94 (0.93 to 0.94)*
AOSW-inferred MSE	*0.92 (0.92 to 0.92)*	*1.00*	*0.99 (0.98 to 0.99)*
AOSW norm MSE	*0.94 (0.93 to 0.94)*	*0.99 (0.98 to 0.99)*	*1.00*

Table 6.3. Genetic correlations for the traits Autorefraction MSE, AOSW-inferred MSE, and AOSW norm MSE. 95% confidence intervals are shown in brackets.

6.3.5 Effect Size Correlations for Most Strongly Associated Markers

Figure 6.9 demonstrates the correlations of regression (beta) coefficients for genetic variants from the GWAS analyses for Autorefraction MSE and AOSW-inferred MSE. There was a significant amount of noise seen when all variants were plotted, however as the variants were filtered by their significance of association to the phenotype (i.e. P value), a positive correlative was observed (note that P value filtering was applied to both traits (i.e. the same P value threshold was applied to both traits simultaneously when determining the correlation). Table 6.4 shows the correlation coefficients for these different levels of significance.

The correlation of beta coefficients for the traits Autorefraction MSE and AOSW norm MSE are shown in Figure 6.10. A similar pattern of positive correlation was observed after filtering by the degree of association. The correlation coefficients are shown in Table 6.5.

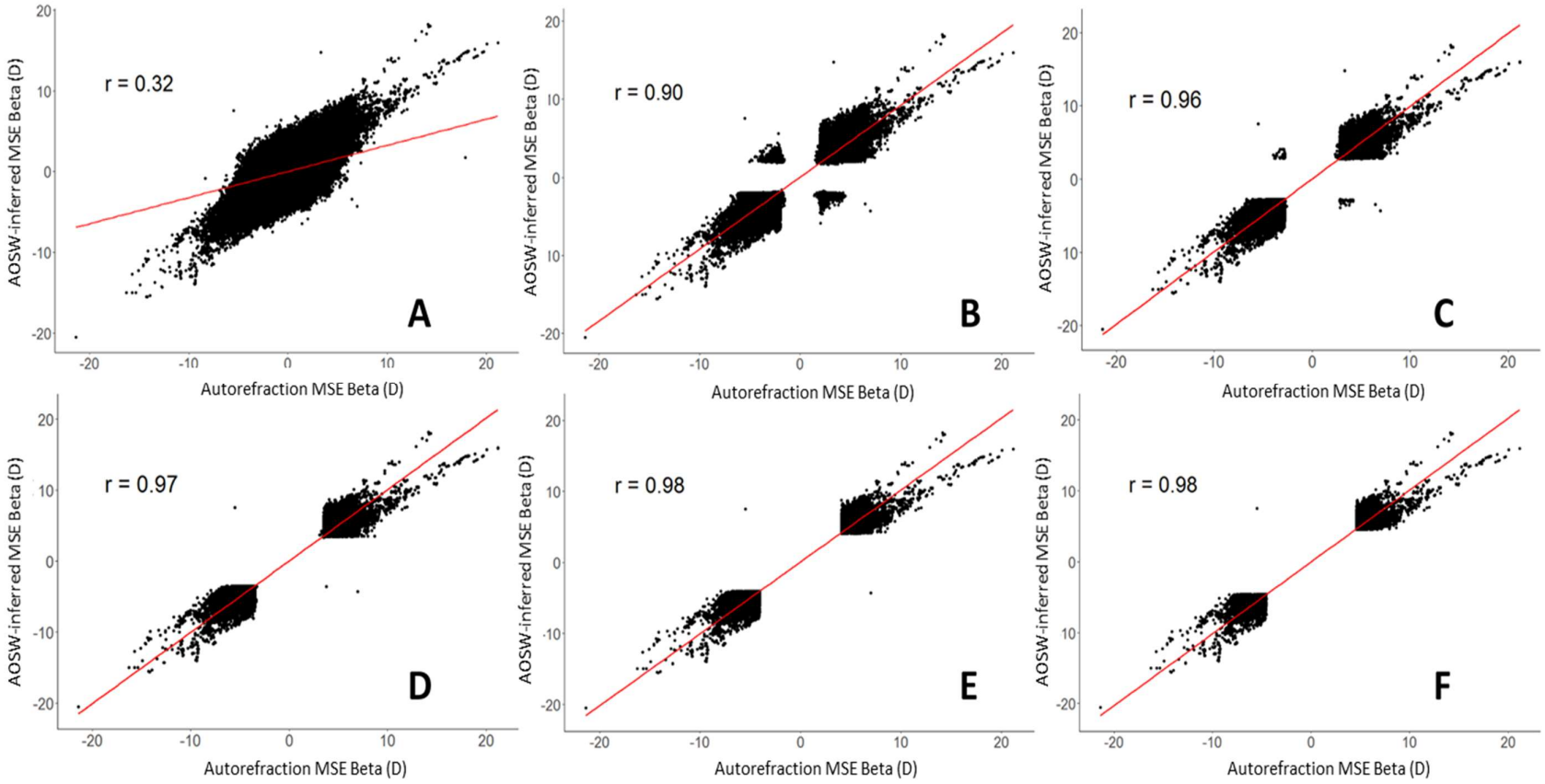


Figure 6.9. Graphs demonstrating the correlation of effect sizes for variants in the GWAS for Autorefraction MSE and AOSW-inferred MSE. Panels A to F indicate the P value filter: (A) no filter, (B) P < 0.5, (C) P < 0.05, (D) P < 0.005, (E) P < 0.0005, (F) P < 0.00005, applied in both the Autorefraction MSE and AOSW-inferred MSE data.

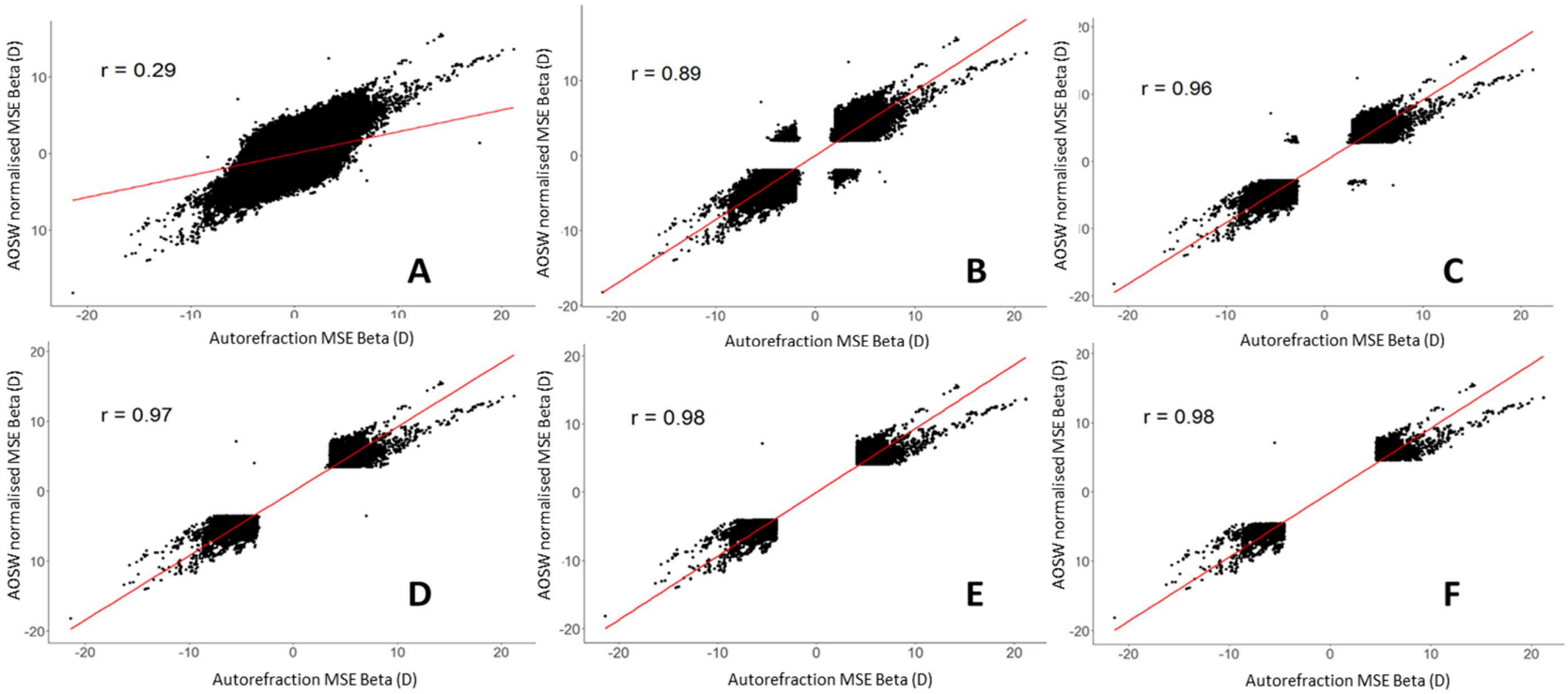


Figure 6.10. Graphs demonstrating the correlation of effect sizes found in the genetic variants of the GWAS summary statistics for Autorefraction MSE and AOSW norm MSE. Panels A to F indicate the P value filter: (A) no filter, (B) P < 0.5, (C) P < 0.05, (D) P < 0.005, (E) P < 0.0005, (F) P < 0.00005 applied in both the Autorefraction MSE and AOSW-inferred MSE data.

Filtered Significance Level	Number of Genetic Variants remaining	Effect Size Correlation (95% CI)
None (all variants)	9,767,769	0.317 (0.316 – 0.317)
$5x10^{-1}$	265,812	0.898 (0.897 – 0.899)
$5x10^{-2}$	79,498	0.964 (0.963 – 0.965)
$5x10^{-3}$	40,213	0.972 (0.972 – 0.973)
$5x10^{-4}$	25,733	0.976 (0.975 – 0.977)
$5x10^{-5}$	17,146	0.979 (0.978 – 0.980)

Table 6.4. Effect size correlations between variants in the Autorefraction MSE and AOSW-inferred MSE GWAS summary statistics. Variants were filtered by GWAS P value.

Filtered Significance Level	Number of Genetic Variants remaining	Effect Size Correlation (95% CI)
None (all variants)	9,767,769	0.292 (0.293 – 0.294)
$5x10^{-1}$	232,539	0.889 (0.889 – 0.890)
$5x10^{-2}$	76,544	0.962 (0.961 – 0.962)
$5x10^{-3}$	43,284	0.968 (0.967 – 0.969)
$5x10^{-4}$	32,123	0.976 (0.975 – 0.977)
$5x10^{-5}$	18,256	0.976 (0.976 – 0.977)

Table 6.5. Effect size correlations between variants in the Autorefraction MSE and AOSW normalised MSE GWAS summary statistics. Variants were filtered by GWAS P value.

6.4 Discussion

The analyses reported in this chapter attempted to make use of the 287,448 individuals of European ancestry from the UK Biobank cohort who did not have refractive data available. The results demonstrated that AOSW, age and gender could be used to infer refractive error. Both AOSW and age had non-linear relationships with refractive error. (Note that it was not possible to determine if the association with age (or YOB) was due to changes that occur to an individual's refractive error with age, or a cohort effect in the overall population, such as a shift towards more a negative refraction in younger generations). The optimised prediction model employing these 3 variables yielded an 'AOSW-inferred MSE' phenotype and an 'AOSW norm MSE' phenotype that explained approximately 30% and 45% of the variance of refractive error, respectively. Moreover, the strong genetic correlation ($r_g > 0.92$) identified for AOSW-inferred MSE and AOSW norm MSE compared to Autorefraction MSE gives support to the idea that these two

imputed phenotypes are reasonable surrogates for autorefraction-measured refractive error. This conclusion is further supported by the positive correlation in effect sizes demonstrated in Figure 6.9 and Figure 6.10.

The R^2 of 0.30 for the linear regression model of AOSW-inferred MSE demonstrates that the factors used in the regression model explain approximately 1/3 of the inter-individual variation in refractive error. This can be extrapolated to mean that the AOSW-inferred MSE trait is ~30% as effective at explaining the variation in refractive error as autorefraction data would be, and therefore having a sample of participants with known AOSW-inferred MSE of approximately 3x the size of the Autorefraction MSE sample would provide an approximate doubling of the effective sample size. For example, in terms of GWAS statistical power, 286,515 participants with known AOSW-inferred MSE would be equivalent to 95,505 participants with known Autorefraction MSE. This is similar to the number of AOSW-inferred MSE participants actually available (N = 287,448). Therefore, by performing a GWAS in the AOSW-inferred MSE sample, it is possible to increase the effective sample size in detecting refractive error variants by 2-fold compared to the level when only using participants with known Autorefraction MSE.

In a similar manner, the R^2 of AOSW-norm MSE was higher at 0.45, indicating that it explained a greater proportion of the variance in refractive error than the AOSW-inferred MSE trait. Therefore, as a better predictor of refractive error, a GWAS for this trait should further improve the ability to detect loci for refractive error compared to a GWAS for AOSW-inferred MSE.

Both GWAS Miami plots demonstrated similarly located loci for the predicted phenotypes and Autorefraction MSE. However, the GWAS analyses for the predicted phenotypes yielded much reduced log P values. This occurs as a result of the imprecision in predicting refractive error using the prediction method above.

Genetic correlations calculated with LDscore regression yielded positive values for all pairwise trait comparisons, with the highest genetic correlation being between AOSW-inferred MSE and AOSW norm MSE. This is likely to be an inflated result as the two phenotypes are related to each via a simple, non-linear transform.

The next largest genetic correlation was for AOSW norm MSE vs. Autorefraction MSE, at r_g = 0.94, which was higher than that for non-transformed AOSW-inferred MSE vs. Autorefraction MSE (r_g = 0.92). Moreover, since the confidence intervals of these correlations do not overlap, this implied a statistically significant difference for the two genetic correlations (this was confirmed using the correlation difference test in R; P < 0.001). The improvement in genetic correlation after transforming the phenotype to a normal distribution may be due to the distribution of AOSW norm MSE being more similar to the distribution of Autorefraction MSE (see Figure 6.4 and Figure 6.6). Moreover, this result suggests that in future analyses designed to combine GWAS summary statistics from different traits, AOSW norm MSE may perform better for prediction when combined with Autorefraction MSE than AOSW-inferred MSE. However, broadly speaking it appears that all 3 traits demonstrate a strong positive correlation, meaning that all trait summary statistics (other than the 2 inferred phenotypes) could be combined to increase the effective sample size used for analysis, and potentially obtain a more accurate prediction estimate for refractive error and myopia.

The trend of increasing positive correlation between predicted phenotypes and Autorefraction MSE when filtering on significance indicates that loci associated with refractive error were being identified accurately in the GWAS for the inferred phenotypes. Nevertheless, this also suggests that if we are to combined GWAS summary statistics from different phenotypes to create genetic risk scores, it may be beneficial to restrict variants to those that are statistically significant. However this theory is contraindicated by the literature, which suggests that limiting the number of SNPs used for genetic risk score calculation limits predictive accuracy (Vilhjálmsson et al. 2015) (see Section 1.3.10). Therefore, it may be beneficial to include as many variants as possible in future analyses, i.e. both significant and non-significantly-associated variants, despite the potential to add noise.

In conclusion, an optimised multivariable model was used to impute a new phenotype termed ‘AOSW-inferred MSE’, from the variables AOSW, Age and Sex. (An inverse-normal transformed phenotype, ‘AOSW norm MSE’ was also imputed). The AOSW-inferred MSE phenotype provided approximately one third of the effective statistical power to detect genetic variants compared to using participants’ autorefraction-

measured refractive error, meaning that a samples with the imputed phenotypes would need to be 3x larger in order to provide a doubling of the *effective* sample size for a GWAS.

Combining GWAS summary statistics for Autorefraction MSE and AOSW-inferred MSE (or Autorefraction MSE and AOSW norm MSE) should result in an effective GWAS sample size larger than the largest GWAS meta-analysis for refractive error reported to date (160,420 participants) (Tedja et al. 2018). Such a combined GWAS sample would have the potential to improve genetic risk score prediction accuracy compared to previous reports (Pickrell et al. 2016; Tedja et al. 2018).

7 Prediction of Refractive Error Using Correlated Traits

7.1 Introduction

In Chapter 4, an analysis was conducted to determine whether knowing a child's number of myopic parents (i.e. having 0, 1 or 2 myopic parents) or a genetic risk score for refractive error would have better accuracy at predicting the child's refractive error. The analysis used a genetic risk score derived from 149 genetic variants that had been identified in a GWAS for Autorefraction MSE in 95,505 European UK Biobank participants and a GWAS carried out by the CREAM consortium. The results demonstrated that the number of myopic parents was a better predictor of refractive error and incident myopia in children than the genetic risk score.

In this chapter, a similar analysis was conducted, but with the aim of creating an improved genetic risk score for refractive error. Here, the key question was whether the variance in refractive error explained by the new (improved) genetic risk score would reach a level sufficient to achieve clinical utility.

The limitations of the genetic risk score used in Chapter 4 – and the approaches used to validate it and assess its clinical applicability – were considered, with the aim to overcome the previous limitations and make improvements where possible. Firstly, the independent 'validation sample' of participants used for testing the accuracy of the genetic risk score was changed from children to adults. An adult validation sample would be expected to overcome the potential inaccurate estimation of refractive error at earlier ages when the phenotype has not developed completely and stabilized. In other words, using a validation sample comprised of children could conceivably lead to an incorrect categorisation of refractive group, e.g. an individual from the ALSPAC cohort who became myopic in their late teens, beyond the age of 15, would have been miscategorised as non-myopic. Additionally, the children from ALSPAC had their refractive error measured using non-cycloplegic autorefraction (Williams et al. 2008a). There is known to be some discrepancy between the results obtained with and without cycloplegia prior to autorefraction (Williams et al. 2008a; Northstone et al. 2013). This body of literature has identified inaccuracy in non-cycloplegic autorefraction when testing children or young adults (Krantz et al. 2010; Mimouni et al. 2016; Sankaridurg et al. 2017). Using an adult validation sample was expected to overcome this issue and

improve the reliability in the refractive measurements obtained, due to the fact that non-cycloplegic autorefraction is the gold-standard in adults (Sanfilippo et al. 2014) and may therefore improve the accuracy of the prediction found using a genetic risk score.

Secondly, previous studies have suggested that using a genetic risk score composed only of variants associated with the trait-of-interest at genome-wide statistical significance level usually leads to inferior levels of genetic prediction than if additional variants not reaching this significance level had been included (Dudbridge 2013; Marquez-Luna et al. 2017; Lee et al. 2018). This phenomenon has been suggested to be a result of using under-powered GWAS analyses due to insufficient sample sizes, which may have led many truly trait-associated SNPs to show only suggestive levels of association. The 149 genetic variants used in the previous genetic risk score analysis may therefore not have been the optimal choice of variants. Indeed, for biobank-scale datasets, it has been suggested that all available variants should be included in a genetic risk score to improve accuracy (Khera et al. 2018). Therefore, in the current analysis, all of the genetic variants available in the GWAS summary statistics from UK Biobank were included when creating the genetic risk scores.

Thirdly, another approach to increase the power and accuracy of a genetic risk score is to combine information from multiple GWAS analyses (Tedja et al. 2018; Turley et al. 2018). The MTAG (Multi-Trait Analysis of GWAS) software package allows GWAS summary statistics for two or more related (i.e. genetically correlated) traits to be combined together. This approach may be ideally suited to combining the GWAS summary statistics for refractive error, AOSW-inferred MSE, and even educational attainment (see below). Combining the results from any of these GWAS could potentially improve predictive accuracy (e.g. combining AOSW-inferred MSE, educational attainment, and Autorefraction MSE would increase the effective sample size for our trait-of-interest) and therefore increase the predictive power. This approach is discussed further in the Methods section.

A genetic correlation between educational attainment and refractive error was expected based on the literature suggesting strong evidence for both the correlation of educational attainment and myopia risk (Fan et al. 2016b; Morgan et al. 2017) as well as evidence for the causal role of educational attainment in myopia development (Cuellar-

Partida et al. 2015; Mountjoy et al. 2018). Thus, to increase the effective sample size by including genetically correlated traits, summary statistics from a GWAS for educational attainment performed by Okbay et al. (2016) would be used, which were available through an open access website.

When creating a genetic risk score, the LD (section 1.3.6) of nearby SNPs has to be taken into account. The reason for this can be appreciated through an extreme example: consider two SNPs in perfect LD (i.e. $r^2 = 1$), one of which has a major causal role in myopia while the other has no causal role whatsoever. In a GWAS analysis, the regression coefficient for both SNPs would be the same. Consequently, a genetic risk score derived *without* considering LD would, erroneously, assign equal weight to each SNP, whereas ideally the effect assigned to one of the SNPs should be zero (Vilhjálmsson et al. 2015). Thus, ignoring LD biases the apparent effect size of neighbouring non-causative SNPs, which reduces the accuracy of the effect estimate of adjacent causal SNPs, leading to a reduced genetic risk score accuracy. Here, LDpred software (Section 3.2.5) was used to account for LD, and improve the accuracy of genetic prediction.

The analyses in this experiment were used to investigate the following hypotheses:

1. That combining GWAS summary statistics of genetically correlated traits with the use of MTAG would improve the genetic prediction accuracy for refractive error compared to the GWAS summary statistics used for each trait in isolation.

2. Combining GWAS summary statistics using MTAG would improve the genetic prediction accuracy for refractive error compared to conventional inverse-variance weighted meta-analysis.

3. That the accuracy of using a genetic risk score calculated from a larger effective sample size to predict refractive error and myopia in adults would reach the level required for clinical utility.

7.2 Methods

7.2.1 Participant Selection

The selection of UK Biobank participants and the resultant sample sizes have been discussed in Section 2.1.3.

The following summary statistics were used in the analyses described in this chapter:

1) GWAS summary statistics for the phenotype 'Autorefraction MSE' obtained from a sample of N = 95,505 UK Biobank participants of European ancestry (see Section 5.2).

2) GWAS summary statistics for the phenotype 'AOSW-inferred MSE' in a sample of N = 287,448 UK Biobank participants of European ancestry (see Section 6.2.1 for details of how the AOSW-inferred MSE phenotype was derived, and Section 6.2.3 for how the GWAS summary statistics were obtained).

3) GWAS summary statistics for the phenotype 'AOSW norm MSE' in a sample of N = 287,448 UK Biobank participants of European ancestry (see Section 6.2.2.1 for details of how the AOSW-inferred normalised MSE phenotype was derived, and Section 6.2.3 for how the GWAS summary statistics were obtained). Note that the GWAS for this phenotype included exactly the same sample of participants as the GWAS for AOSW-inferred MSE. However the phenotype differs in that AOSW norm MSE was derived by rank-based inverse-normal transformation of the AOSW-inferred MSE, such that the distribution of the trait was normal prior to GWAS analysis. As the AOSW-inferred MSE and AOSW norm MSE groups have 100% overlap in their participants, their GWAS summary statistics were not combined in subsequent analyses.

4) GWAS summary statistics for the phenotype 'EduYears' (years spent in full time education) for N = 328,917 participants of European ancestry from Okbay et al. (2016). These summary statistics were downloaded from the Social Science Genetic Association Consortium (SSGAC) website (https://www.thessgac.org/data). All SSGAC participants had been questioned about their level of education at the age of 30 or above, when it would be expected that most participants would have attained their highest level of full-time education. The SSGAC GWAS meta-analysis included participants distributed across many different educational systems and countries of birth; therefore, the outcome variable 'EduYears' was derived by Okbay et al. using the 1997 International Standard Classification of Education from the United Nations to allow for comparative and collective analysis. It should be noted that

approximately 40,000 individuals from the interim UK Biobank release were included in the Okbay et al. GWAS analysis, which will have led to a degree of overlap with the above-mentioned samples. However, MTAG meta-analysis (see below) can allow for some overlap in the samples used, and therefore these summary statistics were deemed suitable for inclusion in the current study. However, conventional meta-analysis methods require the use of non-overlapping samples, and therefore for this reason the GWAS summary statistics for the EduYears phenotype were excluded from the conventional meta-analysis.

As described in Section 3.1.4, all GWAS summary statistics were filtered to remove any potential low quality data or missing/incorrect genetic data.

Numbers of participants in the samples for analysis for all traits, is shown in Table 7.1.

Traits and Trait Combinations	Number of Participants	Use of MTAG	Use of METAL
Autorefraction MSE	95,505	N	N
AOSW-inferred MSE	287,448	N	N
AOSW norm MSE	287,448	N	N
EduYears	328,917	N	N
Autorefraction & AOSW-inferred MSE	383,067	Y	Y
Autorefraction & AOSW norm MSE	383,067	Y	Y
Autorefraction MSE & EduYears	424,536	Y	N
Autorefraction MSE, AOSW-inferred MSE, EduYears	711,984	Y	N
Autorefraction MSE, AOSW norm MSE, EduYears	711,984	Y	N

Table 7.1. A table of all traits and trait combinations with their respective sample sizes used to create a genetic risk score. Whether MTAG was used is also listed. Trait combinations used for conventional inverse-variance weighted meta-analysis with METAL are also listed.

7.2.2 Independent Validation Sample

Mothers from the ALSPAC cohort were chosen as the independent validation sample (Section 2.2). There were N = 1,516 ALSPAC mothers of European genetic ancestry who had autorefraction data available and who passed quality control filtering as described in Section 3.1.4.

7.2.3 Genetic Correlation Assessment

Genetic correlations for Autorefraction MSE, AOSW-inferred MSE, and AOSW norm MSE were calculated in the previous chapter (Section 6.3.4). However, additional genetic correlations involving EduYears were calculated here, using LD score regression (Section 3.2.3).

7.2.4 Multi-Trait Analysis of Genome Wide Association Summary Statistics (MTAG)

MTAG software was used to combine genetically correlated traits together for downstream analysis. Details of MTAG analysis are provided in Section 3.2.4. In order for data files to be read by MTAG, beta values from GWAS summary statistics were transformed to Z scores, and column headers were changed to the MTAG input file defaults. As MTAG gives separate output files for each of the traits used in the analysis (i.e. it gives different weighted output files for each trait of interest), the MTAG output file for the desired trait (Autorefraction MSE) was taken and used in all successive analyses. These results were then processed with LDpred (Section 3.2.5; see below) to account for LD.

7.2.5 Conventional Inverse Variance Weighted Meta-Analysis Using METAL

Two combinations: Autorefraction MSE and AOSW-inferred MSE, and Autorefraction MSE AOSW norm MSE, were combined using the inverse variance weighted method implemented in METAL (Willer et al. 2010). This was performed using the MTAG input files with beta coefficient transformed Z scores. The Z scores were then reconverted to beta values post analysis. The output from the meta-analysis was then processed with LDpred (see below). The accuracy of the genetic risk scores derived using METAL meta-analysis were then compared to those derived from MTAG.

7.2.6 LDpred

All analyses were run using the 'LDpred-inf' option i.e. assuming an infinitesimal model (Section 3.2.5). LDpred analysis was carried out for each trait separately (Autorefraction MSE, AOSW-inferred MSE, AOSW norm MSE, and EduYears) as well as for the MTAG and METAL combined trait combinations (Table 7.1).

The first step of LDpred requires a reference panel to calculate LD patterns. As a sample of at least 2000 unrelated individuals has been recommended as the reference panel for LDpred (Vilhjálmsson et al. 2015), 2500 unrelated female participants of European ancestry were chosen at random from the UK Biobank sample for use as the LD reference panel. Only females were used in the creation of the panel in order to enable the X chromosome to be modelled using the same approach as for autosomes, i.e. individuals would have genotypes of 0, 1 or 2 rather than simply 0 or 1.

For the second step, LDpred uses a Gibbs sampler, which is a mathematical process of determining probabilities of variables using their relative comparison to other variables which can be determined. Gibbs sampler algorithms allow the estimation of unknown factors (in this case the genotype-phenotype association for specific pairs of genetic variants), through a multivariable probability distribution when directly sampling the raw genotype data is not possible, which occurs when only GWAS summary statistics and estimates of the population-side LD from a reference panel are available. LDpred uses the results from this Gibbs sampler to alter posterior effects from GWAS summary statistics by accommodating LD patterns identified in the first step. The software also provides options to adjust settings within the algorithm for this second step, such as the number of iterations or 'loops' the algorithm performs to determine the mean relative probability of genetic variants being present together.

For this experiment, the 'LD radius' parameter was selected as 1000 base pairs. This value corresponds to the number of SNPs on either side of the test SNP for which LDpred adjusts LD, i.e. the larger the LD radius, the larger the region that is adjusted for LD. An LD radius of 1000 was chosen for logistical reasons to minimise computational time, but still maintain the best predictive ability. Furthermore, Vilhjalmsson et al. (2015) recommend at least 60 iterations of the Gibbs sampler during LDpred. Here, 200 iterations were used, since the use of fewer iterations sometimes led to failure of the

Gibbs sample to converge for one or more chromosomes (which led to a poor performance of the generated genetic risk score).

In the final step, the variant weights obtained using LDpred were used to create a genetic risk score (Section 3.2.6). A 'raw' genetic risk score was also created using the non-LDpred-adjusted variants weights (i.e. weights obtained directly from MTAG or METAL, without adjustment for LD) for comparison.

7.2.7 Assessment of Refractive Error and Myopia Risk

Each genetic risk score (both raw and weighted) derived from all phenotypes and phenotype combinations, was used as a predictor variable in a regression analysis for the outcome variable autorefraction-measured refractive error in the validation sample of 1,516 ALSPAC mothers. Prediction accuracy was quantified using the R^2 value. To estimate the accuracy of prediction of myopia in the validation sample, the area under the receiver operating characteristic (AUROC) curve was calculated. This was performed for three myopia severity thresholds: any level of myopia (MSE ≤ -0.75D); moderate myopia (MSE ≤ -3.00D); high myopia (MSE ≤ -5.00D). Likewise, to determine if the genetic risk score was able to discriminate the risk of myopia development, the odds ratio for myopia was calculated for participants in the validation sample in the top 25th, 10th, and 5th percentile of the genetic risk score vs. the remaining lower risk participants.

7.3 Results

7.3.1 Genetic Correlations

Genetic correlations calculated using LD score regressions are listed in Table 7.2.

	Autorefraction MSE	**AOSW-inferred MSE**	**AOSW norm MSE**	**EduYears**
Autorefraction MSE	*1.00*	*0.92 (0.92 to 0.92)*	*0.94 (0.94 to 0.94)*	*-0.26 (-0.26 to -0.26)*
AOSW-inferred MSE	*0.92 (0.92-0.92)*	*1.00*	*0.99 (0.99-1.00)*	*-0.35 (-0.35 to -0.35)*
AOSW norm MSE	*0.94 (0.94 to 0.94)*	*0.99 (0.99-0.99)*	*1.00*	*-0.33 (-0.33 to -0.33)*
EduYears	*-0.26 (-0.26 to -0.26)*	*-0.35 (-0.35 to -0.35)*	*-0.33 (-0.33 to -0.33)*	*1.00*

Table 7.2. Genetic correlations between traits. Values in brackets indicate 95% confidence intervals. Adapted from Table 6.3 to also include genetic correlations between ocular phenotype traits and EduYears.

As shown in Table 7.2, negative genetic correlations were observed between EduYears and the refractive error-related traits. These negative genetic correlations arise because a negative refractive error is associated with greater educational attainment. Traits with a positive or negative genetic correlation are equally suitable for combining using MTAG (the alleles for either trait can be switched to obtain beta coefficients of the same magnitude but opposite direction). This 'correction' is performed by MTAG automatically. Therefore, EduYears was used in the combined trait MTAG analyses without manual adjustment of the beta coefficients.

7.3.2 Comparison of Genetic Prediction Between METAL and MTAG

Trait Combination	Prediction accuracy (R^2)	
	METAL meta-analysis	**MTAG meta-analysis**
Autorefraction MSE and AOSW-inferred MSE	7.6% (6.3 –8.9%)	10.8% (7.9 – 13.8%)
Autorefraction MSE and AOSW norm MSE	7.4% (6.1 – 8.7%)	9.7% (7.0 – 12.6%)

Table 7.3. Accuracy in predicting refractive error in an independent validation sample for pairs of refractive error-related traits meta-analysed using either MTAG or METAL. Note that variant weights were adjusted for LD using LDpred after performing the meta-analysis. 95% confidence intervals are shown in brackets.

Table 7.3 presents the prediction accuracy results (R^2 values) for traits meta-analysed using either METAL or MTAG (note that variant weights were adjusted for LD using LDpred prior to use in calculating genetic risk scores). For both pairs of traits examined, the MTAG-derived genetic risk score had a better predictive performance than the METAL-derived genetic risk score (model fit; $P \leq 0.001$ for both).

7.3.3 Accuracy of Genetic Risk Scores in Predicting Refractive Error

Table 7.4 and Figure 7.1 show the prediction accuracy (R^2 values) for genetic risk scores derived using the GWAS summary statistics for each trait separately or combined (using MTAG). Results are presented for both raw and LD-adjusted (LDpred) variant weights.

The model with the best predictive accuracy for refractive error was the model with combined Autorefraction MSE, AOSW-inferred MSE, and EduYears, with an R^2 of 11.2%. Although confidence intervals overlapped for many of these genetic risk scores, the model fit indicates that the addition of information relating to educational attainment improved predictive performance compared to the genetic risk score with the second highest accuracy, combined Autorefraction MSE and AOSW-inferred MSE (R^2 11.2% vs. R^2 10.8%, P = 0.005).

Trait/Trait Combination	R^2 raw effects	R^2 weighted effects
Autorefraction MSE	5.0% (2.8-7.1%)	7.1% (4.7 – 9.7%)
AOSW-inferred MSE	3.7% (1.9-5.6%)	6.9% (4.5 – 9.4%)
AOSW norm MSE	2.8% (1.2-4.4%)	5.2% (3.1 – 7.5%)
EduYears	0.01% (0.0-0.1%)	0.14% (0.0 – 0.6%)
Autorefraction MSE & AOSW-inferred MSE	5.8% (3.5-8.1%)	10.8% (7.9 – 13.8%)
Autorefraction MSE & AOSW normalised MSE	5.3% (3.1-7.5%)	9.7% (7.0 – 12.6%)
Autorefraction MSE & EduYears	4.8% (2.7-6.9%)	7.9% (5.4 – 10.6%)
Autorefraction MSE, AOS-inferred MSE & EduYears	5.8% (3.8-8.0%)	11.2% (8.3 – 14.2%)
Autorefraction MSE, AOSW normalised MSE & EduYears	5.2% (3.1-7.4%)	10.1% (7.3 – 13.0%)

Table 7.4 Raw and weighted genetic risk score effects of all traits and combined traits. R^2 values have been provided in percentage format. 95% confidence intervals have been provided in brackets.

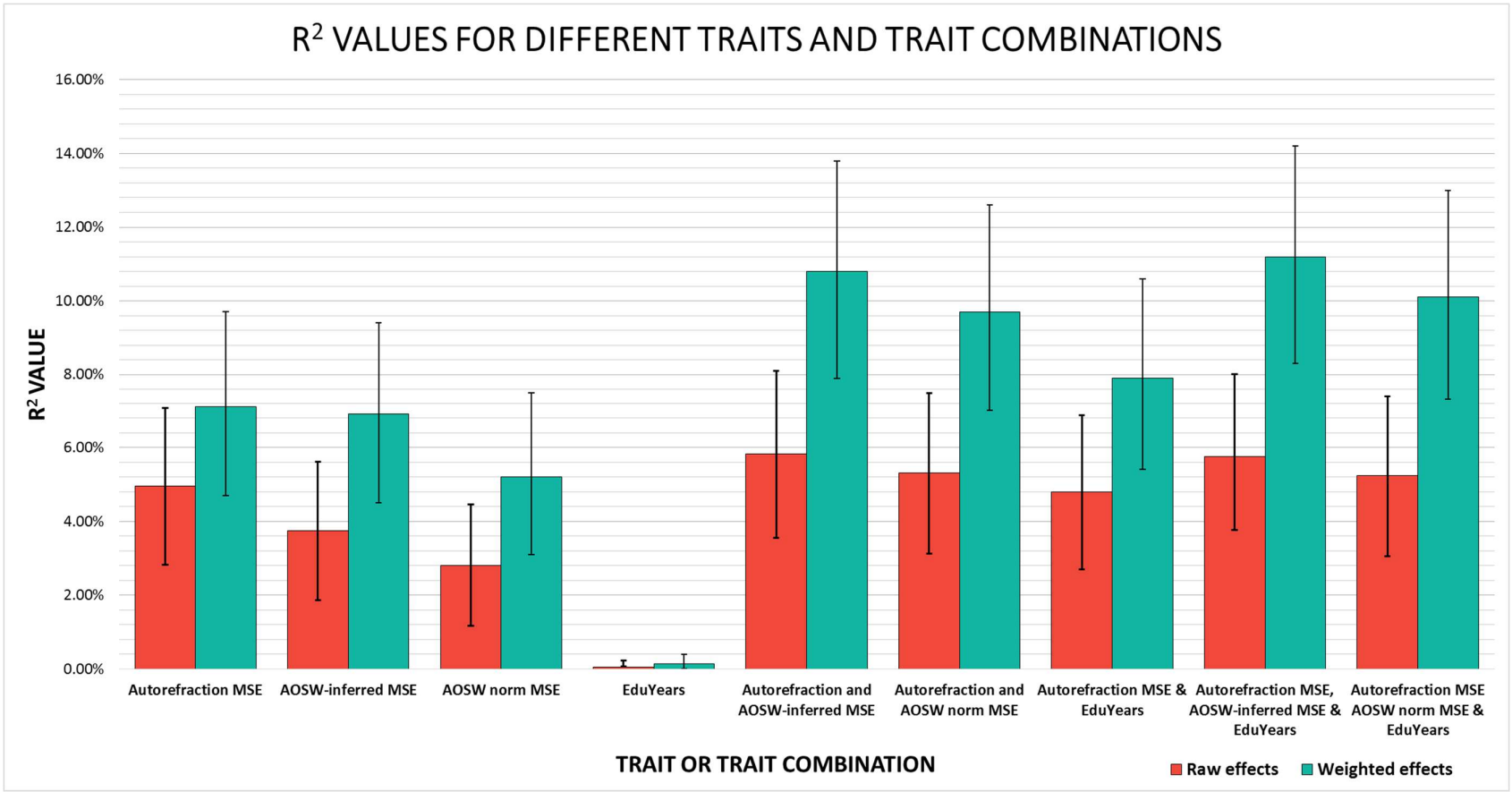
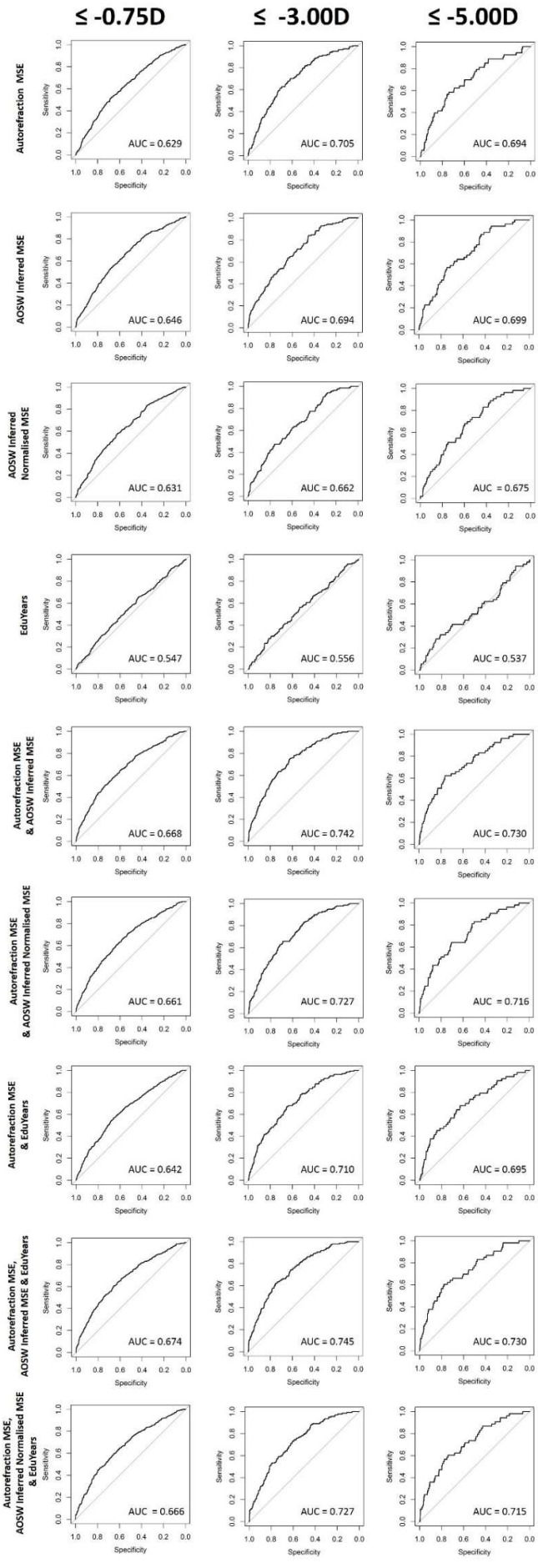


Figure 7.1. Accuracy in prediction of refractive error using a genetic risk score derived from a range of single or combined GWAS summary statistics. Raw effects correspond to variant weightings not adjusted for LD, Weighted effects correspond to variant weighting adjusted for LD using LDpred. Error bars indicate 95% confidence intervals.

7.3.4 Predicting Myopia Status using Genetic Risk Scores

The accuracy of the LD-adjusted weighted genetic risk scores in predicting myopia was examined, using the AUROC to quantify potential clinical utility. Three separate thresholds for classifying myopia severity were considered: any myopia (≤-0.75D), moderate myopia (≤-3.00D), and high myopia (≤-5.00D). The results are shown in Figure 7.2 and Table 7.5. There was a general trend for the various models to yield AUROC point estimates of higher accuracy for the trait 'moderate myopia' compared to 'any myopia' or 'high myopia'. However, 95% confidence intervals often over-lapped, indicating a lack of statistical evidence to support a clear difference between predictive performance for the different myopia severities. The combined model created using GWAS summary statistics for Autorefraction MSE, AOSW-inferred MSE, and EduYears had the best predictive ability, consistent with the results for predicting refractive error (Table 7.5). Using a bootstrap ROC test comparison measure between this model, and the model that included Autorefraction MSE and AOSW-inferred MSE (which provided the next best prediction values) found a slightly improved AUROC for myopia (0.668 vs. 0.674, P = 0.02) but not the AUROC for moderate or high myopia (0.745 vs. 0.742, P = 0.61, and 0.730 vs. 0.730, P = 0.98, respectively).

Figure 7.2. (Overleaf) ROC curves quantifying the accuracy of predicting myopia of varying degrees of severity using a genetic score created using GWAS summary statistics for the specified trait or trait combinations. The AUROC is indicated at the bottom right of each ROC curve panel.


≤ -0.75D
≤ -3.00D
≤ -5.00D
Autorefraction MSE
AOSW Inferred MSE
AOSW Inferred Normalised MSE
EduYears
Autorefraction MSE & AOSW Inferred MSE
Autorefraction MSE & AOSW Inferred Normalised MSE
Autorefraction MSE & EduYears
Autorefraction MSE, AOSW Inferred MSE & EduYears
Autorefraction MSE, AOSW Inferred Normalised MSE & EduYears
AUC = 0.629
AUC = 0.705
AUC = 0.694
AUC = 0.646
AUC = 0.694
AUC = 0.699
AUC = 0.631
AUC = 0.662
AUC = 0.675
AUC = 0.547
AUC = 0.556
AUC = 0.537
AUC = 0.668
AUC = 0.742
AUC = 0.730
AUC = 0.661
AUC = 0.727
AUC = 0.716
AUC = 0.642
AUC = 0.710
AUC = 0.695
AUC = 0.674
AUC = 0.745
AUC = 0.730
AUC = 0.666
AUC = 0.727
AUC = 0.715
Sensitivity
Specificity

Myopia severity threshold	Autorefraction MSE	AOSW-inferred MSE	AOSW norm MSE	EduYears	Autorefraction & AOSW-inferred MSE	Autorefraction & AOSW norm MSE	Autorefraction MSE & EduYears	Autorefraction MSE, AOSW-inferred MSE & EduYears	Autorefraction MSE, AOSW norm MSE & EduYears
≤-0.75D	0.629 (0.600-0.659)	0.646 (0.617-0.676)	0.631 (0.601-0.612)	0.547 (0.515-0.579)	0.668 (0.639-0.698)	0.661 (0.631-0.690)	0.642 (0.612-0.672)	0.674 (0.645-0.704)	0.666 (0.637-0.696)
≤-3.00D	0.705 (0.661-0.749)	0.694 (0.650-0.737)	0.662 (0.617-0.707)	0.556 (0.505-0.607)	0.742 (0.702-0.784)	0.727 (0.685-0.768)	0.710 (0.666-0.754)	0.745 (0.704-0.786)	0.727 (0.685-0.769)
≤-5.00D	0.694 (0.620-0.768)	0.699 (0.632-0.766)	0.675 (0.607-0.743)	0.537 (0.452-0.622)	0.730 (0.661-0.799)	0.716 (0.649-0.786)	0.695 (0.620-0.770)	0.730 (0.660-0.801)	0.715 (0.644-0.785)

Table 7.5. Accuracy of predicting myopia of varying degrees of severity using a genetic score. Values show AUROC (with 95% CI). Each column gives the results of a genetic risk score model created using GWAS summary statistics for the specified trait or combination of traits. Genetic risk score variant weights were adjusted for LD using LDpred.

7.3.5 Assessment of Clinical Utility

The results above suggested that the most accurate genetic risk model for predicting refractive error and myopia was the model created by combining GWAS summary statistics for Autorefraction MSE, AOSW-inferred MSE, and EduYears. This model yielded an accuracy R^2 = 11.2% and an AUROC of 0.67 for refractive error and myopia prediction, respectively. Therefore, this model was further examined to evaluate the clinical utility of using genetic prediction for detecting individuals at high risk of developing myopia.

Odds ratios for myopia (of severity level: any, moderate, or high) were calculated for individuals with genetic risk scores in the top 25%, 10%, and 5% of the sample compared to the remaining 75%, 90% and 95% as the reference sample. A visual representation of this analysis is shown in Figure 7.3.

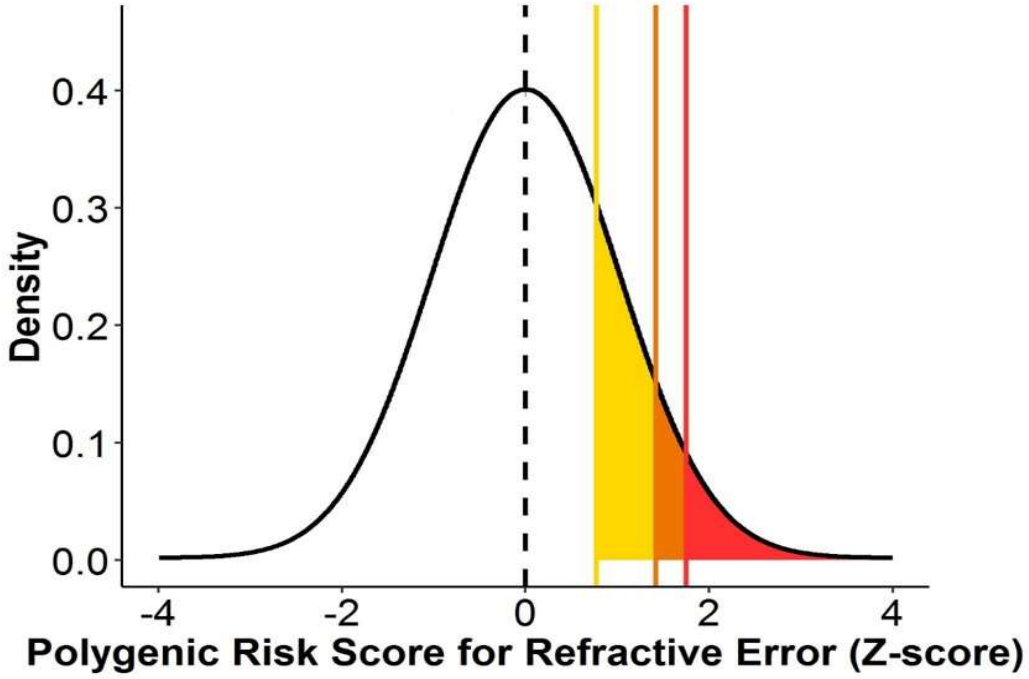


Figure 7.3. Selection of participants with genetic risk scores in the top 25%, 10% or 5% of the distribution. The genetic risk scores have been standardised to aid interpretation. A more positive Z score value indicates a higher genetic risk of myopia. The shaded regions correspond to the top 25th, 10th, and 5th percentile of the population, which were examined as the high risk groups.

The results show an increased risk of being myopic when being in the top 25% or higher genetic risk categories than would be expected by chance. Full results are shown in Table 7.6.

There appears to be a gradual increase in the odds of being myopic for individuals with increasingly higher levels of genetic risk i.e. the risk of being myopic for individuals in the top 25%, 10%, and 5% of the genetic risk distribution was associated with a steady increase in the risk of myopia (from 3x , to 4.6x, and 4.9x, respectively). This accumulative risk pattern was also be seen for moderate and high myopia, with the greatest risk of high myopia (≤ -5.00D) found in the top 5% of individuals, who were at 6.5x increased risk compared to the remaining 95% of the sample. Furthermore, the level of risk appeared to be greater when looking at individuals at higher risk percentiles compared to lower ones. For example, the risk of being highly myopic for the top 25%, 10%, and 5% of individuals was 4.6x, 5.4x and 6.5x, respectively. Overall, it appears as though stratification on the basis of genetic risk was predictive of both the chance of developing myopia and the degree of myopia an individual is likely to attain.

Trait	Risk group	Reference group	Odds ratio (95% CI)	*P*-value
Myopia ≤ -0.75D	Top 25%	Remaining 75%	3.06 (2.40 – 3.91)	1.75×10^{-19}
	Top 10%	Remaining 90%	3.47 (2.43 – 4.91)	9.70×10^{-13}
	Top 5%	Remaining 95%	4.57 (2.84 – 7.51)	7.11×10^{-10}
Myopia ≤ -3.00D	Top 25%	Remaining 75%	4.66 (3.06 – 7.03)	3.93×10^{-13}
	Top 10%	Remaining 90%	4.89 (3.41 – 7.06)	8.14×10^{-18}
	Top 5%	Remaining 95%	5.42 (3.17 – 9.03)	1.95×10^{-10}
Myopia ≤ -5.00D	Top 25%	Remaining 75%	4.90 (2.81 – 8.72)	3.22×10^{-8}
	Top 10%	Remaining 90%	6.11 (3.36 – 10.87)	1.20×10^{-9}
	Top 5%	Remaining 95%	6.50 (3.14 – 12.48)	1.37×10^{-7}

Table 7.6. Odds ratios for having myopia of at least ≤-0.75D, ≤-3.00D, and ≤-5.00D for individuals categorised as being at high risk according to their genetic risk score (being in the top 25%, 10% or 5% of the distribution). Odd ratios were calculated by comparing those in the high risk group to the remainder of the population (reference group).

7.4 Discussion

In this Chapter, GWAS summary statistics for the traits Autorefraction MSE, AOSW-inferred MSE, AOSW norm MSE, and EduYears were used to derive a series of genetic risk scores for predicting refractive error in an independent validation sample. Multiple traits were combined using MTAG, and used to derive genetic risk scores. LD between nearby genetic markers was adjusted for using LDpred. Prediction accuracy varied from approximately 0.1% to 11.2% at the maximum. Individuals with genetic risk scores in the top 10% were approximately 6-fold more likely to develop myopia of ≤-5.00D, which suggests the model could have some clinical utility.

The genetic correlation between the refractive error phenotypes was discussed in Chapter 6. Genetic correlations between the refractive phenotypes and educational attainment showed a moderate negative correlation. The values of -0.35 and -0.33 for the genetic correlations of AOSW-inferred MSE and AOSW norm MSE with EduYears, respectively, were numerically higher than the genetic correlation with Autorefraction MSE at -0.26. The 95% confidence intervals for these correlations did not overlap, suggesting that genetic variants associated with educational attainment may be more comparable to genetic variants for the age of first wearing glasses/need for ocular correction, rather than autorefraction-measured refractive error.

The findings in relation to the comparison of METAL and MTAG indicate that genetic risk scores derived from MTAG meta-analysis gave better predictions of refractive error than those from METAL when using the same data (R^2 = 7.6% vs. 10.8%, and R^2 = 7.4% vs. 9.7%; model fit, P < 0.001 for both). There was an improvement in prediction accuracy after combining information for Autorefraction MSE along with the traits AOSW-inferred MSE or AOSW norm MSE, irrespective of whether the meta-analysis was performed with MTAG or METAL. The improvement in prediction accuracy after combining Autorefraction MSE and AOSW-inferred MSE with MTAG was 3.7% (7.1%, 95% CI 4.7 – 9.7% vs. 10.8%, 95% CI 7.9 – 13.8%; P = 0.001). The improvement in prediction accuracy after combining the two phenotypes with METAL was 0.5% (7.1%, 95% CI 4.7 – 9.7% vs. 7.6% (95% CI 6.3 –8.9%); P = 0.004). The results for AOSW norm MSE were similar to those for AOSW-inferred MSE, with their combination using both MTAG and METAL showing an improved prediction accuracy compared to Autorefraction MSE alone, and

a higher accuracy found when combining the traits through MTAG rather than using METAL.

Lee et al. (2018) conducted a study in which educational attainment and the correlated traits of self-reported maths ability, cognitive performance, and highest maths class level taken, were combined using MTAG. The authors stated that they had combined these traits because they all had a genetic correlation of 0.50 with their trait-of-interest, and therefore it was permissible to use MTAG to combine the traits together. My results suggest that a lower threshold genetic correlation of 0.26 is an acceptable genetic correlation level for a trait to be included in an MTAG meta-analysis, as the inclusion of EduYears to the genetic risk score model derived only from Autorefraction MSE summary statistics improved the model fit (7.9% vs. 7.1%; model fit, P = 0.0002).

Table 7.7 lists studies that have used MTAG to combine traits, and the minimum genetic correlation between the primary trait-of-interest and the other traits.

Study	Number of Traits	Minimum Genetic Correlation
Day et al. (2018)	3	0.69
Hill et al. (2018)	2	0.70
Lee et al. (2018)	3	0.51

Table 7.7. Studies that have used MTAG and the minimum genetic correlation between the primary trait of interest and the other traits.

Both raw and weighted genetic risk scores were evaluated (Table 7.5). The prediction accuracy (R^2 value) for the weighted risk score was up to 50% higher than that for the raw score, illustrating the importance of accounting for LD. Therefore, all subsequent analyses used the LDpred-weighted genetic risk scores.

The maximum prediction accuracy achieved with a genetic risk score was of R^2 = 11.2%, which was larger than the previous highest figure of 7.8% reported for a genetic risk score derived from a meta-analysis of GWAS summary statistics for refractive error and age of onset of myopia (N = 160,420 participants) carried out by the CREAM consortium and 23andMe (Tedja et al., 2018).

The results in Section 6.2.1 indicated that a GWAS for AOSW-inferred MSE (and AOSW norm MSE) in 286,515 participants should have similar statistical power as a GWAS for Autorefraction MSE in 95,505 participants. (This was because the variance in refractive error explained by the AOSW-inferred MSE phenotype was $R^2 \approx 0.3$, and there were approximately 3 times more participants with data for AOSW-inferred MSE compared to Autorefraction MSE). This would imply they should have a similar 'effective' sample size. This was supported in the genetic risk score analyses: the predictive accuracy of Autorefraction MSE and AOSW-inferred MSE were very similar (R^2 = 7.1% vs. 6.9%, respectively; model fit, P = 0.45). However, AOSW norm MSE performed significantly worse in refractive error prediction than AOSW-inferred MSE (R^2 = 7.1% vs. 5.2%; model fit, P = 0.0001). This result was unexpected, and the reason for the inferior performance of AOSW norm MSE is currently unclear.

As mentioned above, combining Autorefraction MSE and AOSW-inferred MSE produced a genetic risk score that performed better than using Autorefraction MSE alone (10.8% vs. 7.1%; likelihood ratio test, P < 0.0001). Nevertheless, the improvement in accuracy was not double, despite the doubling of the 'effective' sample size. This is likely due to the reduced improvement in predictive accuracy that is expected as the GWAS sample size increases. Further evidence of this phenomenon of 'diminishing returns' with increasing sample size was noted for EduYears, too (see below). It should be noted that this is also likely to be due to the lower genetic correlation between EduYears and Autorefraction MSE, and poorer overall predictive performance of EduYears when used in isolation (genetic risk score derived from EduYears alone: R^2 = 0.014%).

Nonetheless, combining GWAS summary statistics for EduYears with the refractive error-related traits *did* marginally improve the fit of predictive models (as mentioned above). For example, the inclusion of EduYears in the MTAG meta-analysis improved the fit of prediction models when combined with: Autorefraction MSE (7.1% vs. 7.9%, P = 0.0002), Autorefraction and AOSW-inferred MSE combined (10.8% vs. 11.2%, P = 0.005), and Autorefraction and AOSW norm MSE (9.7% vs. 10.1%, P=0.006).

On average, there was a 0.6% increase in accuracy for the weighted genetic risk score after combining information from the EduYears trait. Thus, it can be concluded that

including summary statistics for educational attainment is beneficial for genetic prediction of refractive error, even though the degree of improvement is modest.

Although the inclusion of EduYears when deriving genetic risk scores always enhanced the prediction of *refractive error*, this was not always the case for the prediction of *myopia*. The inclusion of GWAS summary statistics for EduYears along with those for Autorefraction MSE did significantly improve prediction of 'any myopia': AUROC = 0.629 vs. 0.642 (P = 0.01). Conversely, the inclusion of GWAS summary statistics for EduYears did *not* improve prediction of moderate or high myopia (P = 0.5 and P = 0.9, respectively). This pattern was also seen when GWAS summary statistics for EduYears were combined with those for Autorefraction MSE and AOSW-inferred MSE. Prediction of 'any myopia' improved (AUROC = 0.668 vs. 0.674, P = 0.02), but there was no improvement found for predicting moderate and high myopia (P > 0.70). This raises the question of whether the inclusion of EduYears in myopia prediction has any real value, particularly as it showed poor prediction accuracy when used alone.

The above findings suggest that the inclusion of EduYears improves the accuracy with which the level of refractive error can be predicted (i.e. higher R^2 value), and also the accuracy of predicting low levels of myopia, but not in predicting more severe levels of myopia. It may be that individuals who are already at high genetic risk of severe myopia are not at an appreciably higher risk if they are genetically predisposed to educational attainment, but that individuals who are not as genetically predisposed to myopia may become so due to increased educational attainment. Although the ability to draw strong conclusions is limited, based on the data from this chapter the results are consistent with the theory that education is influencing the distribution of refractive error in the population currently. Nonetheless it is highly probable that if the sample size of Autorefraction MSE was to increase, that this finding of an improvement in prediction of 'any myopia' following the inclusion of an EduYears may be contradicted. Thus, the benefit of incorporating information about EduYears may reflect the limited sample size of the Autorefraction MSE GWAS, consistent with the lack of improved prediction by inclusion of EduYears once AOSW-inferred MSE had already been combined with Autorefraction MSE. However, the use of a GWAS for EduYears in a much larger sample of participants may be beneficial in predicting myopia, should larger GWAS samples for refractive error not be forthcoming in the future.

There was a significant difference in the prediction of 'any myopia' when comparing models created using GWAS summary statistics for Autorefraction MSE alone vs. AOSW-inferred MSE alone (0.629 vs. 0.646, ROC bootstrap test, P = 0.01). However, this was not the case for predicting moderate or high myopia.

The pattern of results regarding the prediction of myopia had an overarching correspondence to the pattern of results for prediction of refractive error, e.g. both displayed optimal performance for the 3 trait MTAG model incorporating Autorefraction MSE, AOSW-inferred MSE and EduYears. There was also a common trend for models created using AOSW norm MSE to perform more poorly than those incorporating (the unadjusted) AOSW-inferred MSE. For this reason, the former model was used in assessing the clinical utility of genetic prediction of myopia.

As discussed by Torkamani et al. (2018), a polygenic risk score that achieves a sufficient level of prediction will allow for stratification of the population into different sub risk categories. My analysis confirmed that categorisation of individuals into groups with different risks of myopia is possible. The results shown in Table 7.6 demonstrate that an increasingly high level of genetic risk is associated with an increased risk of myopia as well as with a more severe level of myopia. For example, the genetic risk score could be used to divide a population into a group at low risk of myopia development, comprising of 75% of the starting sample, and a high risk group comprising of the remaining 25% of the sample who are at 3-times increased risk of developing myopia. This high risk sample could be subdivided further, to select 5% or 10% of the population who are at a further-increased risk of moderate or high myopia.

Thus, in theory genetic risk scores could be used by eye care providers to tailor patients to different management options, such as more regular screening of at-risk individuals. Children in the upper 5-25% of the genetic risk score model could be advised to attend more regular screening check-ups and spend more time outdoors. The top 5% of the genetic risk score distribution - who are at a 6.5 times increased risk of high myopia - may benefit strongly from prophylactic time outdoors, as well as other optical and pharmaceutical myopia intervention methods for delaying myopia onset, should any become established.

The best level of myopia prediction achieved for the genetic risk score, AUC = 0.75 for moderate myopia, is still not as good as that found by other studies. The CLEERE study (Zadnik et al. 2015) calculated an AUC of 0.87 for the development of myopia when using a cycloplegic autorefraction of +0.75D or less at the age of 6 years. However, genetic prediction still holds some value, as it can be done before the age of 6 years (e.g. from birth) before any biological mechanisms in leading to myopia development or 'pre-myopia' may have started. It is currently uncertain whether the genetic risk score could be combined with the cycloplegic autorefraction of children at the age of 6 to significantly improve predictive efficacy. Current work suggests not (Chen et al., 2019), but it should be noted that the study by Chen et al. was performed on children between the ages of 7-15 years old, many of whom were already myopic when attending their clinic visits.

GWAS sample size has been argued to be the largest factor limiting in the accuracy of genetic prediction (Dudbridge 2013). Identifying non-additive effects and including them in the genetic risk score model may also help improve accuracy. Nonetheless, as discussed previously, these approaches will be limited in their ability to give drastic improvements in genetic prediction (according to quantitative genetics theory). Currently, SNP-heritability estimates for refractive error put the upper limit of prediction accuracy at 39% (Shah et al. 2018). The analyses here resulted in a predictive accuracy of 11.2%, i.e. approximately 29% of this upper limit.

In conclusion, the results obtained in this chapter showed that using the trait combination of Autorefraction MSE, AOSW-inferred MSE, and EduYears (incorporating GWAS summary statistics data for a combined sample size of N = 711,984) yielded a genetic risk score with the best prediction accuracy; namely, an R^2 value of 11.2% when testing in an independent European sample. This result is an improvement over the 7.8% accuracy reported by the CREAM consortium and 23andMe (Tedja et al., 2018), and is the most accurate genetic prediction estimate for this phenotype to date. Combining genetically correlated traits ($|r_g| > 0.25$) using MTAG improved the accuracy in predicting refractive error and myopia. The genetic risk score created in this chapter may have some clinical utility for detecting children aged < 6 years old at risk of myopia and high myopia. Thus, a personalised medicine approach for myopia management is feasible, at least in theory.

8 Prediction of Refractive Error in Individuals with Non-European Ancestry

8.1 Introduction

In Chapter 7, genetic risk score models were created using a range of GWAS summary statistics either on their own and or combined through MTAG. The accuracy of these genetic risk score models to predict refractive error and myopia in an independent sample of ALSPAC mothers with European ancestry was tested. Compared to the previous literature, the accuracy in predicting refractive error and myopia was improved with the new, multi-trait MTAG models. Furthermore, combining summary statistics from genetically correlated traits was found to be beneficial in improving the genetic prediction of refractive error.

However, the results in Chapter 7 were only applicable to individuals of European ancestry. To date, there have been no published reports of the performance of a genetic risk score for predicting myopia in individuals of non-European ancestry, although Marquez-Luna et al. and Perry et al. have assessed the performance of cross-ethnic polygenic prediction for disorders such as diabetes (Marquez-Luna et al. 2017; Perry et al. 2018). The two latter studies showed that genetic prediction in participants of non-European ethnicity was poorer than in Europeans, and that this deficit represents a large limitation of current genetic research being undertaken worldwide. Diabetes differs in prevalence across the world, being more common in non-Caucasian ethnicities (Wild et al. 2004), which suggests that it would be beneficial to investigate genetic factors and prediction in these ethnicities. This issue is also pertinent to myopia; there is a large discrepancy in the prevalence of myopia across the world (as discussed in Section 1.2.3). In general, the prevalence of myopia is higher in East and Southeast Asian countries, and in individuals of Asian ethnicity (Pan et al. 2012; French et al. 2013b; Morgan et al. 2017). Thus, identifying individuals at a higher risk of developing myopia in non-European populations, particularly in those of East Asian ethnicity, would be useful clinically.

In this chapter, the genetic risk score models created in the previous chapter were tested in individuals of various non-European ethnicities. The analysis was focussed on the relative accuracy in predicting refractive error and myopia in different ethnic groups. To do this, the performance of models to predict refractive error and myopia was compared

to those described in Chapter 7 for the European ALSPAC sample (see Methods). It was hypothesised that if the genetic loci influencing refractive error are shared between different ethnicities, then the prediction of refractive error and myopia will be similarly accurate in individuals of European and non-European ancestry.

8.2 Methods

To compare the accuracy of genetic prediction of myopia for individuals of non-European vs. European ancestry, genetic prediction was carried out in UK Biobank participants of non-European ancestry. These results were then compared against the findings reported in Chapter 7 for the N = 1,516 ALSPAC mothers (European ancestry independent sample; Section 7.2.2).

8.2.1 Participant Selection

UK Biobank participants of non-European ancestry with information available for refractive error were studied. Specifically, individuals who self-reported Asian, Chinese, or Black ethnicity were used in this analysis, resulting in samples of 3,651, 455, and 3,368 adults, respectively. As self-reported ethnicity is not always an accurate portrayal of ethnic background (Mersha and Abebe 2015), individuals were clustered to ensure ancestral homogeneity using principal component analysis (PCA; see Section 3.1.2). The mean and standard deviation for each of the first 10 PCs were calculated, separately, for UK Biobank participants whose self-reported ethnicity was either Asian, Chinese, or Black using PC data from Bycroft et al. (Bycroft et al. 2018). Any individual outside of ±10 standard deviations from the mean of any PC value was then excluded from the analyses. This filtering step ensured that any participant whose genetic ancestry did not cluster with other participants of the same self-reported ethnicity category would be excluded.

8.2.2 Genetic Risk Score Modelling

Once the above filters for genetic ancestry were applied, weighted genetic risk scores were calculated for the remaining participants. The weights used were those obtained from the MTAG meta-analysis and LDpred analysis for the 3 GWAS traits Autorefraction MSE, AOSW-inferred MSE and EduYears, as described in Section 7.2.4 and 7.2.6, i.e. the effect size weights were those derived using GWAS summary statistics from European

samples and a European LD reference panel. The "--SCORE" function in PLINK 1.9 (Purcell et al. 2007) was used to calculate genetic risk scores for each participant (Section 3.2.6).

A linear regression of genetic risk score on autorefraction-measured MSE was carried out to calculate the prediction accuracy (R^2 value) in each ethnic group (Asian, Chinese, and Black). Following the protocol adopted in Chapter 7, prediction accuracy was assessed for each of the 9 MTAG-LDpred models. This resulted in 36 different analyses: 9 MTAG-LDpred models x 4 ethnic groups (Asian, Chinese, Black, and European).

The model with the best predictive accuracy for refractive error (estimated by R^2) in each ethnic group was evaluated for its efficacy in predicting myopia development. This was done using the area under the receiver operating characteristic curve (AUROC), for the three thresholds used in Chapter 7, namely: any level of myopia ≤ -0.75D; moderate myopia ≤ -3.00D; and high myopia ≤ -5.00D. Odds ratios for low, moderate and high myopia were calculated for those in the highest 25th, 10th and 5th percentile of genetic risk compared to the remaining sample.

8.3 Results

8.3.1 Participant Filtering

Table 8.1 demonstrates the number of participants removed by the PCA filtering step (shown visually in Figure 8.1).

Self-reported Ethnicity	Participants before PCA filtering	Participants after PCA filtering
Asian	3,599	3,500
Chinese	454	444
Black	3,366	3,132

Table 8.1. Principle component analysis filtering to exclude participants whose genetic ancestry did not cluster with other participants of the same self-reported ethnicity.

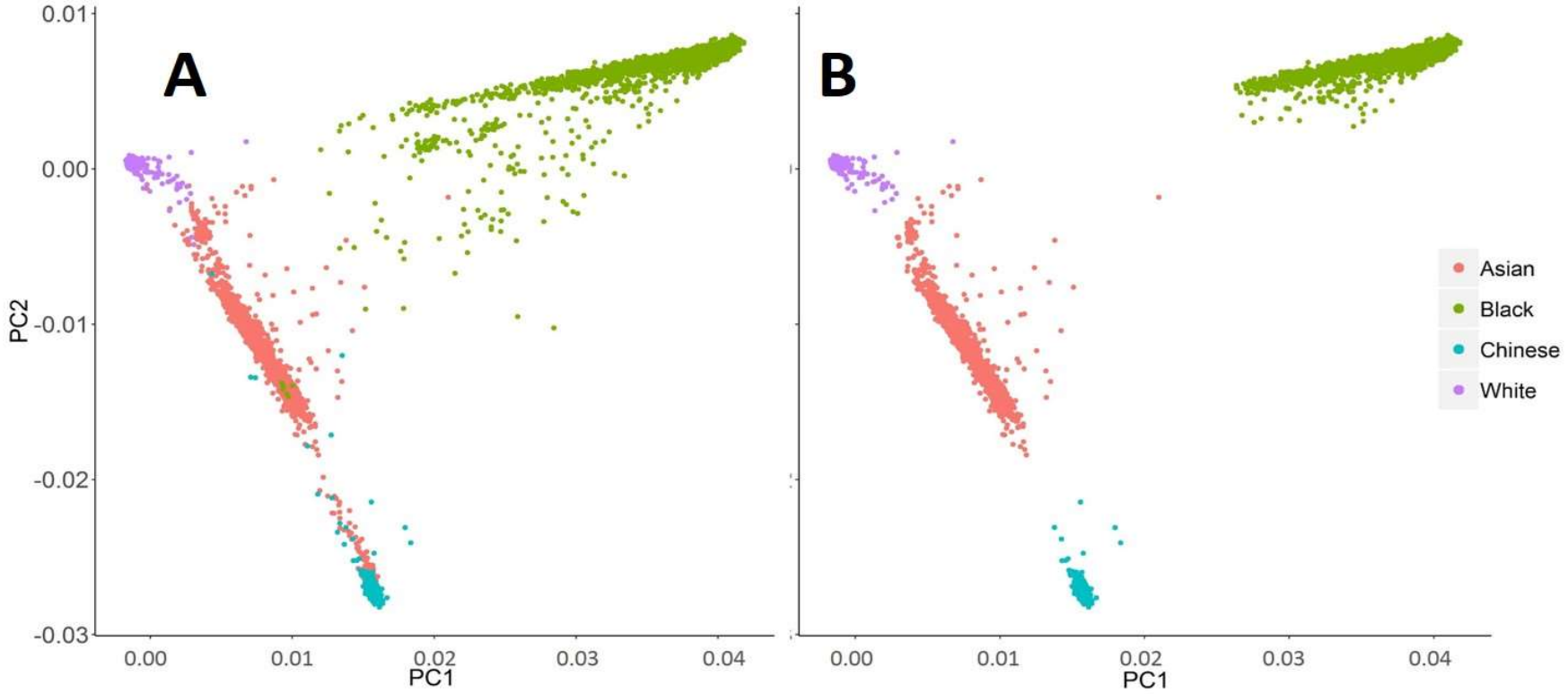


Figure 8.1. Visual depiction of the effect of the PCA filtering step on participants with different self-reported ethnicity. The figure shows a scatterplot of PC1 vs. PC2. Panel A shows the participants before PCA filtering, and Panel B shows participants after filtering. Note that the European sample is comprised of 1,516 randomly selected UK Biobank participants with self-reported White British ethnicity (this group was included in the graphs rather than the ALSPAC Mothers sample since PCs derived from one sample are not directly applicable to another sample i.e. values from the ALSPAC mothers cannot be directly comparable to the samples in UK Biobank).

8.3.2 Refractive Error Prediction

Table 8.2 and Figure 8.2 display the results for the refractive error prediction accuracy.

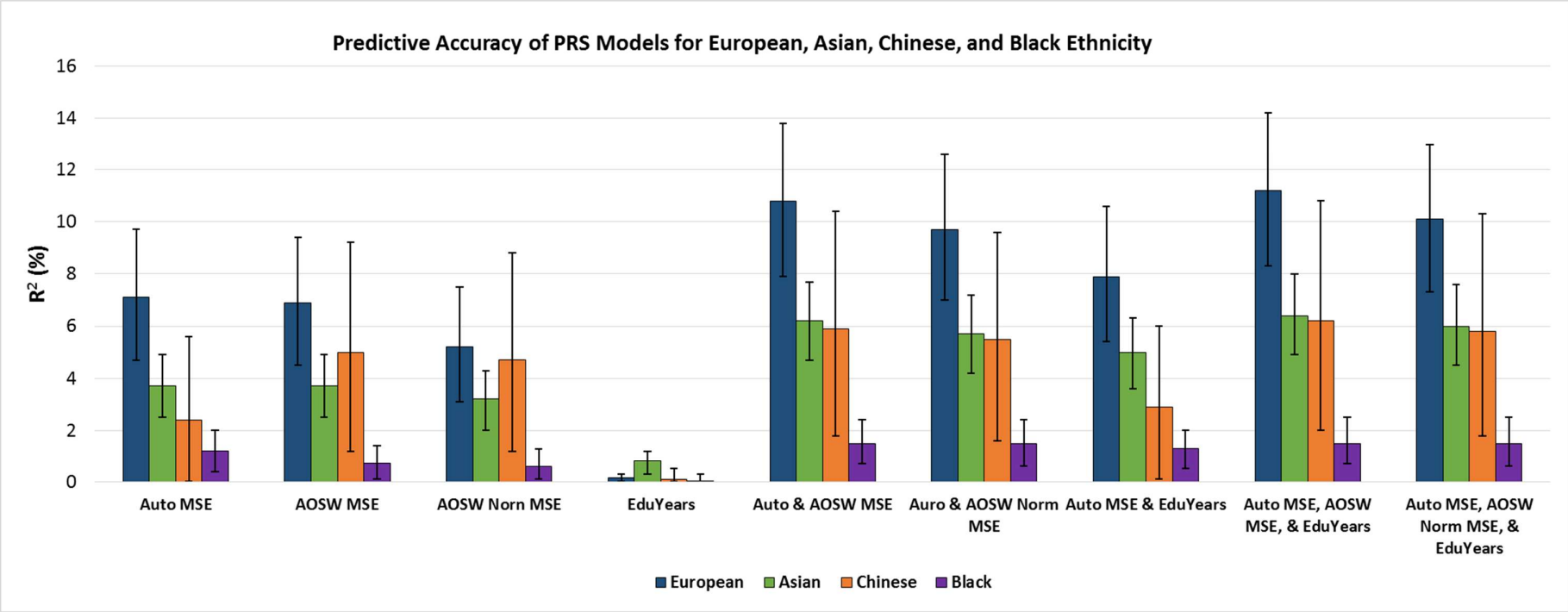


Figure 8.2. Predictive accuracies of 9 different genetic risk score models in individuals of European, Asian, Chinese, and Black ancestry. Error bars indicate 95% confidence intervals.

Trait/Trait Combination	European	Asian	Chinese	Black
Autorefraction MSE	7.1% (4.7 – 9.7%)	3.7% (2.5 – 4.9%)	2.4% (0.0 – 5.6%)	1.2% (0.4 – 2.0%)
AOSW-inferred MSE	6.9% (4.5 – 9.4%)	3.7% (2.4 – 4.9%)	5.0% (1.2 – 9.2%)	0.7% (0.1 – 1.4%)
AOSW norm MSE	5.2% (3.1 – 7.5%)	3.2% (2.0 – 4.3%)	4.7% (1.2 – 8.8%)	0.6% (0.1 – 1.3%)
EduYears	0.14% (0.0 – 0.6%)	0.8% (0.5 – 1.2%)	0.1% (0.0 – 0.4%)	0.0% (0.0 – 0.2%)
Autorefraction MSE & AOSW-inferred MSE	10.8% (7.9 – 13.8%)	6.2% (4.7 – 7.7%)	5.9% (1.8 – 10.4%)	1.5% 0.7 – 2.4%)
Autorefraction MSE & AOSW norm MSE	9.7% (7.0 – 12.6%)	5.7% (4.2 – 7.2%)	5.5% (1.6 – 9.6%)	1.5% (0.6 – 2.4%)
Autorefraction MSE & EduYears	7.9% (5.4 – 10.6%)	5.0% (3.6 – 6.3%)	2.9% (0.0 – 6.0%)	1.3% (0.5 – 2.0%)
Autorefraction MSE, AOSW-inferred MSE & EduYears	11.2% (8.3 – 14.2%)	6.4% (4.9 – 8.0%)	6.2% (2.0 – 10.8%)	1.5% (0.7 – 2.5%)
Autorefraction MSE, AOSW norm & EduYears	10.1% (7.3 – 13.0%)	6.0% (4.5 – 7.6%)	5.8% (1.8 – 10.3%)	1.5% (0.6 – 2.3%)

Table 8.2. Predictive accuracies for the 9 different genetic risk score models for individuals of European, Asian, Chinese, and Black ancestry. R^2 values are displayed as percentages. Values in brackets indicate the 95% confidence intervals.

The genetic risk score derived using GWAS summary statistics for Autorefraction MSE, AOSW-inferred MSE, and EduYears combined performed better for prediction of refractive error than the other models in all 4 ethnic groups. Genetic prediction of refractive error was most accurate in Europeans, intermediate in Chinese and Asians, and least accurate in individuals of Black ethnicity.

8.3.3 Myopia Prediction

The prediction model using GWAS summary statistics for Autorefraction MSE, AOSW-inferred MSE, and EduYears combined was used to estimate the sensitivity and specificity of predicting myopia (quantified as the AUROC). Figure 8.3 displays the ROC curves with the corresponding AUROC values for all four ethnicities using the Autorefraction MSE, AOSW-inferred MSE, and EduYears model.

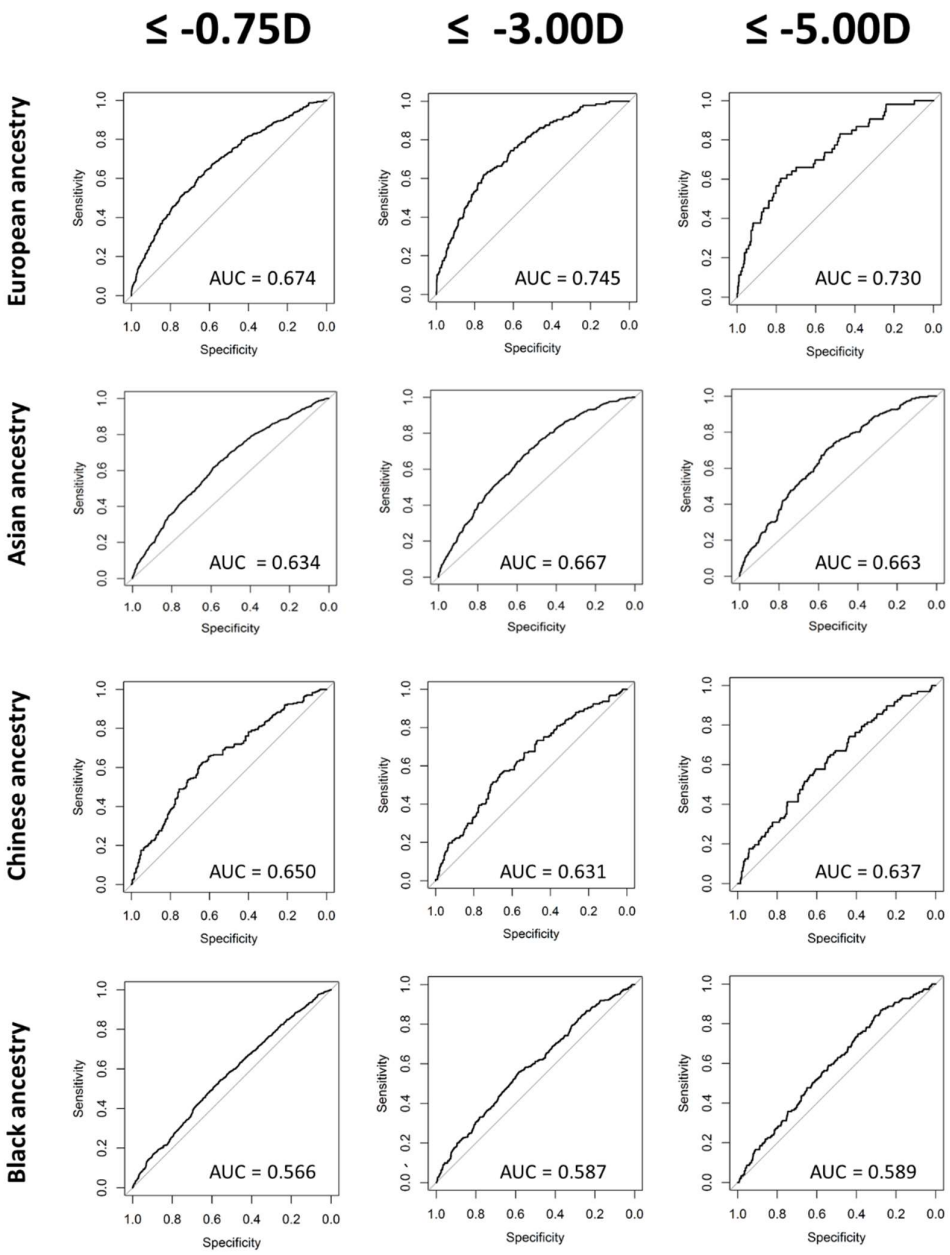


Figure 8.3. The receiver operating characteristic (ROC) curves for predicting any myopia, moderate myopia, and high myopia in the four different ethnic groups using the genetic risk score model derived from Autorefraction MSE, AOSW-inferred MSE, and EduYears. The corresponding area under the curve for each ROC curve in the panel is noted. The results for individuals of European ancestry are those from Chapter 7.

8.3.4 Clinical Applicability

The results demonstrated prediction of refractive error was most accurate with the genetic risk score model derived from GWAS data for the 3 traits: Autorefraction MSE, AOSW-inferred MSE and EduYears. This was the case for all 4 ethnicities. Accuracy (R^2

value) was 11.2%, 6.4%, 6.2%, and 1.5% for European, Asian, Chinese, and Black ancestries, respectively. As the clinical applicability of the genetic risk score in Europeans has been addressed in Section 7.3.5 and the discussion of Chapter 7, it will not be repeated here. However, a similar analysis was performed for the other 3 non-European ancestries: the top 25%, 10%, and 5% of individuals identified as "high risk" were compared to the corresponding remainder of 75%, 90% or 95% of individuals. The odds ratios for predicting any myopia, moderate myopia and high myopia in participants of each ethnicity are presented in Table 8.3, Table 8.4, and Table 8.5.

Trait	**Risk group**	**Reference group**	**Odds ratio (95% CI)**	***P*-value**
Myopia ≤ -0.75D	Top 25%	Remaining 75%	2.60 (1.56 – 3.95)	3.9×10^{-4}
	Top 10%	Remaining 90%	3.51 (1.77 – 5.39)	6.1×10^{-5}
	Top 5%	Remaining 95%	4.02 (1.98 – 6.37)	7.9×10^{-3}
Myopia ≤ -3.00D	Top 25%	Remaining 75%	1.98 (1.36 – 3.15)	4.4×10^{-4}
	Top 10%	Remaining 90%	3.84 (1.99 – 6.63)	3.0×10^{-6}
	Top 5%	Remaining 95%	2.81 (1.16 – 6.02)	2.1×10^{-2}
Myopia ≤ -5.00D	Top 25%	Remaining 75%	1.65 (1.00 – 2.69)	4.5×10^{-2}
	Top 10%	Remaining 90%	2.83 (1.52 – 5.44)	5.1×10^{-2}
	Top 5%	Remaining 95%	3.50 (1.47 – 8.35)	6.4×10^{-2}

Table 8.3. Odds ratios for myopia of at least ≤-0.75D, ≤-3.00D, and ≤-5.00D in Asian individuals categorised as having a high genetic risk score. Odd ratios were calculated by comparing those in the high risk group to the remainder of the population (reference group).

Trait	Risk group	Reference group	Odds ratio (95% CI)	*P*-value
Myopia ≤ -0.75D	Top 25%	Remaining 75%	2.26 (1.45 – 3.60)	$4.2x10^{-4}$
	Top 10%	Remaining 90%	3.86 (1.89 – 8.73)	$4.6x10^{-5}$
	Top 5%	Remaining 95%	3.26 (1.27 – 10.01)	$2.1x10^{-2}$
Myopia ≤ -3.00D	Top 25%	Remaining 75%	1.91 (1.23 – 2.97)	$3.7x10^{-4}$
	Top 10%	Remaining 90%	3.10 (1.66 – 5.93)	$4.4x10^{-5}$
	Top 5%	Remaining 95%	3.02 (1.29 – 7.41)	$1.2x10^{-2}$
Myopia ≤ -5.00D	Top 25%	Remaining 75%	1.67 (1.01 – 2.73)	$4.1x10^{-2}$
	Top 10%	Remaining 90%	2.70 (1.40 – 5.12)	$2.6x10^{-3}$
	Top 5%	Remaining 95%	3.57 (1.50 – 8.42)	$3.4x10^{-3}$

Table 8.4. Odds ratios for myopia of at least ≤-0.75D, ≤-3.00D, and ≤-5.00D in Chinese individuals categorised as having a high genetic risk score. Odd ratios were calculated by comparing those in the high risk group to the remainder of the population (reference group).

Trait	Risk group	Reference group	Odds ratio (95% CI)	*P*-value
Myopia ≤ -0.75D	Top 25%	Remaining 75%	1.52 (1.19 – 1.93)	$8.2x10^{-4}$
	Top 10%	Remaining 90%	1.69 (1.19 – 2.37)	$2.8x10^{-3}$
	Top 5%	Remaining 95%	1.64 (1.02 – 2.60)	$3.9x10^{-2}$
Myopia ≤ -3.00D	Top 25%	Remaining 75%	1.68 (1.22 – 2.30)	$1.4x10^{-3}$
	Top 10%	Remaining 90%	2.05 (1.34 – 3.07)	$6.8x10^{-4}$
	Top 5%	Remaining 95%	1.80 (0.98 – 3.11)	$4.5x10^{-2}$
Myopia ≤ -5.00D	Top 25%	Remaining 75%	1.38 (0.85 – 2.17)	4.210^{-1}
	Top 10%	Remaining 90%	1.56 (0.81 – 2.79)	$1.5x10^{-1}$
	Top 5%	Remaining 95%	1.12 (0.36 – 2.59)	$8.1x10^{-1}$

Table 8.5. Odds ratios for myopia of at least ≤-0.75D, ≤-3.00D, and ≤-5.00D in Black individuals categorised as having a high genetic risk score. Odd ratios were calculated by comparing those in the high risk group to the remainder of the population (reference group).

For all three non-European ethnic groups, those in the upper percentiles of the genetic risk score distribution were at higher risk of myopia compared to the remainder of the population ($P < 0.05$). The trend observed for Europeans (**Error! Reference source not found.**) in which the risk of myopia and the risk of a higher severity level of myopia increased gradually for participants with a higher genetic risk score, were not observed in these non-European samples (this point is discussed in the next section).

For those with black ancestry, there was no significant difference found for having high myopia (≤ -5.00D) between the higher and lower risk groups at any of the top 25th, 10th, and 5th percentiles. Conversely, a significant difference was found between all risk groups for any myopia and moderate myopia. These findings are likely to be due to the model's reduced predictive performance in this ethnic group, limiting the power of genetic prediction to differentiate and stratify individuals.

The ability of the genetic risk score to predict myopia severity level was better in those of Asian and Chinese ancestries. An increased risk of myopia was observed for individuals in the higher percentiles of the genetic risk score, however this was less consistent than in those with European ancestry.

8.4 Discussion

In this experiment, SNP weights derived from European GWAS samples were used to create genetic risk scores for refractive error that were subsequently assessed in 3 non-European ethnic groups. By doing this, it was possible to assess how accurate predictions in participants of Asian, Chinese, and Black ethnicity would be using the best genetic prediction model currently available. The results supported the hypothesis that genetic prediction using information derived from one ethnic population is inferior when applied to a different ethnic group. There was an approximately 44% reduction in the accuracy of genetic prediction of refractive error in Asian and Chinese individuals compared to prediction in European participants (R^2 = 6.2% and 6.4% vs. 11.2%). Moreover, the similarity of the genetic prediction accuracies in Asian and Chinese individuals most likely reflects their closer ancestry to one another than to Europeans (Jorde and Wooding 2004; Tateno et al. 2014; 1000 Genomes Project et al. 2015; Wall 2017). The results also suggested an even poorer predictive ability for participants of Black ethnicity, with an 87% reduction in accuracy (R^2 = 1.5% vs. 11.2%).

The accuracy of prediction in all three non-European participant samples followed the same pattern of improvement found in European participants, with the use of genetic risk score models derived from summary statistics for genetically correlated traits combined using MTAG showing improved accuracy. The 3-trait MTAG model (Autorefraction MSE, AOSW-inferred MSE, and EduYears) gave the best-performing genetic predictions for all three ethnicities, however, it can be observed that the level of improvement when combining genetically correlated traits was markedly lower than in European participants, particularly when combining the GWAS summary statistics for EduYears. Moreover, as per the European participants, using EduYears on its own gave the poorest prediction of refractive error and myopia.

The clearest example of the lack of improvement in predictive accuracy (R^2 value) when including EduYears in the MTAG model was in Black participants. The R^2 value was nearly identical with the use of SNP weights calculated from the 2 traits Autorefraction MSE and AOSW-inferred MSE when compared to that from the 3 traits Autorefraction MSE, AOSW-inferred MSE, and EduYears (increase in $R^2 < 0.1\%$; $P = 0.89$). Thus, there was no evidence that the inclusion of EduYears improved genetic prediction accuracy in the Black ethnicity sample. The addition of other correlated traits (namely AOSW-inferred MSE and AOSW norm MSE) did appear to give an increase in accuracy compared to using a model derived from Autorefraction MSE alone (model fit; $P < 0.04$ for all) in Black participants. However, the improved model fit was not accompanied by an appreciable improvement in prediction accuracy, which remained fixed at approximately $R^2 = 1.5\%$. Thus, overall, the performance of the genetic prediction model in participants of Black ethnicity was very poor ($R^2 = 1.5\%$).

Further evidence for EduYears *not* having a greater role in predicting refractive error can also be seen in other populations (Table 8.2). In Asians, the inclusion of GWAS summary statistics for EduYears in the genetic risk score models did not improve the variance explained over and above that for the model incorporating only Autorefraction MSE and AOSW-inferred MSE ($R^2 = 6.4\%$ vs. 6.2%; $P = 0.21$). Moreover, this was also the case for the Chinese and Black ethnic groups ($R^2 = 6.2\%$ vs. 5.9%; $P = 0.13$ and $R^2 = 1.5\%$ vs. 1.5%; $P = 0.66$, respectively). However, for the Asian and Chinese ethnicities (but not Black participants) the combination of EduYears and Autorefraction MSE did improve the genetic prediction of refractive error compared to Autorefraction MSE alone (Model fit,

$P < 0.05$ for both). Therefore, this would suggest that in assessing Asian and Chinese individuals for myopia risk using genetic information (and Europeans, as discussed in the previous chapter), the inclusion of GWAS summary statistics for educational attainment may improve accuracy - if the education GWAS sample is large enough.

Accordingly, while the model using Autorefraction MSE, AOSW-inferred MSE, and EduYears did provide some discriminating ability for the 10-25% of Black participants most at risk of any myopia, moderate myopia, and high myopia (OR $\approx$ 1.5), performance was deemed not sufficient for clinical utility. This was due to the limitation of the model to differentiate different levels of risk accurately and to explain the variance in refractive error. These results could also be partly explained by the smaller proportion of participants with myopia in the Black ethnic group compared to the other samples (prevalence of any myopia = 27% in the Black sample, compared to 54%, 42%, and 31% for Chinese, Asian, and European participants, respectively). It may also be reasonable to assume that another reason for the limited success in genetic prediction for Black ancestry may be due to having the most diverse genetic profile (Jorde and Wooding 2004; 1000 Genomes Project et al. 2015). As genetic diversity is inversely correlated with LD, controlling for LD patterns may have caused poorer tagging of causal variants in the GWAS lead SNPs for the Black sample. Thus, the process of controlling for LD using a different reference sample may be increasingly detrimental for genetic prediction in black ancestry, compared to other ancestries.

The performance of genetic prediction in the Asian and Chinese samples provided more promise. Interestingly, EduYears showed better prediction accuracy in Asian participants than in the other 3 ethnicities (0.8% vs. 0.1%). It has been suggested (as discussed in 1.2.3) that the increase in myopia prevalence in many East Asian countries has been due to the impact of education (Morgan et al. 2017), and that countries that have a culture of putting great importance on educational attainment are also the countries that are demonstrating a marked increase in myopia prevalence. The higher proportion of the variance of refractive error explained via genetic preposition to education in Asians vs. non-Asians is consistent with this hypothesis. It should be noted however, that the EduYears-based genetic prediction result for Chinese individuals does not support the hypothesis: the R^2 for the EduYears-only model was low (0.1%) despite the expectation that Chinese individuals would be exposed to similarly high levels of education as the

Asian group. Furthermore, the R^2 95% confidence intervals for the Chinese and Asians overlap, and are very close between Asians and Europeans, suggesting the relatively high EduYears-based genetic prediction result in Asians compared to all other ethnicities may be a false positive finding for this sample. Notably, due to the relatively smaller sample of participants with self-reported Chinese ethnicity, the genetic prediction estimate for the EduYears-only model in Chinese had wide confidence intervals. Therefore, the inconsistency of the EduYears-based genetic prediction results in Asian and Chinese individuals means that it is hard to draw any conclusion as to whether EduYears does explain more of the variation in refractive error in ethnicities with stronger educational emphasis, or whether this is a spurious finding.

The limited (44%) drop in accuracy of genetic prediction in Asian and Chinese individuals compared to Europeans suggests there are shared genetic loci for refractive error in Europeans, Asians, and Chinese individuals. The higher variance explained using the European derived GWAS summary statistics in Asian and Chinese vs. Black individuals is likely to reflect their more similar respective genetic ancestry in comparison to Europeans (Gabriel et al. 2002; Keinan et al. 2007; Tateno et al. 2014; 1000 Genomes Project et al. 2015; Wall 2017). With this level of accuracy, the genetic risk score model may provide some utility for prediction of myopia in individuals of Asian and Chinese ethnicities. The model showed modest performance to detect individuals at increased risk of myopia for those with a genetic risk score in the top 25%, 10%, and 5% percentile compared to the rest of the sample. Specifically, individuals in the high genetic risk score categories had a 3-fold to 4-fold increased risk of any, moderate or high myopia for the best-performing models. However, these models had wide confidence internals, with minimal evidence for improved risk prediction for more severe levels of myopia. For both the Asian and Chinese ethnicities, there was a numeric trend of a higher point estimate of the risk of high myopia (≤ -5.00D) in those in the top 25% vs 10% and 10% vs. 5% genetic risk score percentile. However, the 95% confidence intervals for these estimates overlapped. Such a trend was less evident for any myopia (≤ -0.75D) and moderate myopia (≤ -3.00D). This is likely to be a limitation of the accuracy of the model, suggesting it is able to distinguish individuals with an increased risk, but not able to differentiate them further.

Torkamani et al. (2018) have argued that genetic risk scores could be used for different purposes, some of which would require more accurate and detailed risk stratification, e.g. selecting individuals who would benefit from a therapeutic intervention. It is likely that the genetic risk score model developed here is not accurate enough to select individuals who would benefit from an intervention that also carried a risk. This is because a risk-benefit analysis would be less favourable for intervention given that the risk of myopia with a high genetic risk score is not only uncertain, but there would be no predictive information regarding the degree of myopia that an individual would go on to develop. However, more benign interventions such as recommending increased time outdoors or more regular sight tests would increase the positive aspects of a cost-benefit analysis (i.e. relatively low risk and cheap to implement), thus meaning that our genetic risk score may be of some merit for a benign intervention.

Genetic prediction accuracy was higher in the European sample than in the Asian and Chinese samples. This finding could be due to one, or a combination, of the following factors:

1. Not all genetic risk variants contributing to refractive error development are shared between Europeans and Asians/Chinese individuals
2. The effect size of causal variants differs between Europeans and Asians/Chinese individuals
3. Genetic variants included in the genetic risk score tag causal variants poorly in Asians/Chinese individuals compared to Europeans, due to differences in LD structure
4. There are differences in the heritability of refractive error between Europeans and Asians/Chinese individuals.

In relation to the first point, a 'POPCORN' analysis identified a high genetic correlation (r_g = 0.79) for refractive error in Asians vs. Europeans (Tedja et al. 2018). This demonstrated that - at least for commonly occurring genetic variants - the same variants cause susceptibility to refractive error in the two ethnic groups. In relation to the second point, the POPCORN analysis described above also suggests that genetic variant effect sizes are similar in Asians vs. Europeans. Therefore, the first and second points are unlikely to be major reasons for the lower accuracy of genetic prediction in Asians.

As regards the third point, it is well-known that patterns of LD vary between Europeans and Asians (Evans and Cardon 2005). Furthermore, the true causal variants contributing to variation in refractive error are unknown, such that the variants that are strongly weighted by LDpred will not always be causal variants (which implies that causal variants will sometimes be assigned weights that are too low). Accordingly, the differing patterns of LD between Europeans and Asians could be a major contributor to the poor performance of genetic risk scores in Asians. In regards to the fourth point, as discussed in Section 1.3.7, heritability estimates are dependent on the sample being assessed and their exposure to environmental risk factors. Therefore, it may be that heritability is different in European and Asian populations. If refractive error was less heritable in Asians, for example due to a relatively greater contribution of lifestyle risk factors, then this ancestry group would have a lower SNP-heritability and would yield lower prediction accuracy for genetic risk scores.

To improve the accuracy of the genetic risk score prediction and achieve clinical relevance, the results could be improved by collecting large cohorts of Asian, Chinese, and Black participants with genetic and refractive data, in order to conduct GWAS analyses in each ancestry. This would be expensive and time consuming. However, at present, alternative approaches for improving genetic risk scores have not been forthcoming.

In conclusion, there was an approximate efficacy of ~56% ($R^2 \approx 6.3\%$ vs. 11.2%) prediction accuracy when using a genetic risk score model derived from European participants to predict refractive error in individuals of Asian or Chinese ethnicity. This indicates substantial sharing of genetic loci for refractive error between these ancestries. The genetic risk score model did not perform as well for those with self-reported Black ethnicity ($R^2 \approx 1.5\%$). Although the accuracy of genetic prediction in Asian and Chinese participants was relatively low ($R^2 \approx 6.3\%$ on average), the prediction models did have some ability to discriminate those at risk of myopia development (e.g. the odds ratio for myopia was 3x higher in those with a genetic risk score in the top 25% vs. the remainder). Arguably, this may provide sufficient discriminative performance to advise lifestyle changes or more regular screening for those at a high genetic risk of myopia.

9 Discussion, Conclusions, and Future Work

9.1 General Discussion

The aim of the research performed in this thesis was to explore the concept of genetic prediction for myopia and refractive error, making use of recently released genetic data for the UK Biobank cohort. Initially, 149 GWAS loci identified in a GWAS carried out by the CREAM consortium (Tedja et al., 2018) were used to create a genetic risk score, which was evaluated in an independent dataset of children from the ALSPAC cohort. The genetic risk score was compared to another predictive factor that would potentially be available for children from birth onwards, namely the child's number of myopic parents. It was found that the genetic risk score and the number of myopic parents were partially independent in their capacity to predict refractive error and myopia. Nevertheless, the genetic risk score derived from the 149 genome wide significant GWAS variants showed limited accuracy ($R^2 \approx 1.1$ - 2.6%), and did not perform as well as knowledge of the child's number of myopic parents ($R^2 \approx 3.0$ - 4.8%). Because of this, a GWAS was then conducted for autorefraction measured refractive error in 95,505 UK Biobank participants. Genetic variants identified at a genome-wide significance level ($P < 5 \times 10^{-8}$) were compared to those published by the CREAM consortium (Tedja et al., 2018) and Pickrell et al. (2016). This demonstrated some evidence of replication; 100 SNPs replicated in the CREAM analysis, and all the available 50 SNPs from Pickrell et al. were replicated. 49 SNPs did not show genome wide significant replication, most likely due to the smaller sample size used, leading to reduced power. This therefore meant that using only the summary statistics for the GWAS of Autorefraction MSE would likely show limited improvement in predicting refractive error accurately.

Steps were taken in order to improve the accuracy of a genetic risk score compared to the model derived from the top 149 GWAS variants, or that likely obtained from the Autorefraction MSE GWAS. Specifically, these steps were: increasing the effective sample size by incorporating information from a GWAS for AOSW-inferred refractive error, adjusting the genetic variant weightings (GWAS beta coefficients) by incorporating information from a genetically correlated trait (educational attainment), increasing the number of genetic variants used to derive the genetic risk score (from 149 to approximately 1.1 million), and using an independent validation sample of adults

rather than children (since adults would have established phenotypes, and be unaffected by the lack of cycloplegia). This led to a better prediction accuracy model for refractive error: R^2 = 11.2%, and an AUC of 0.75 for the prediction of moderate myopia (≤-3.00D). The marked improvement in accuracy of this model demonstrated the ability of the genetic risk score to stratify individuals with different levels of risk.

Moreover, an investigation was carried out into the accuracy with which the 1.1 million variant genetic risk score derived from GWAS analyses in Europeans could predict refractive error in individuals of non-European ancestry. This is the first time such an approach has been examined with regards to refractive error. The results indicated a reduction in predictive accuracy for non-European ancestries. For example, accuracy in Asian and Chinese participants was R^2 = 6.4% and 6.2%, respectively, a 55-57% efficacy compared to that in European individuals. This result still suggested the genetic risk score would have some clinical utility in identifying individuals with an elevated risk of myopia, although not as effectively as those with European ancestry. In contrast, the analysis showed that the existing genetic risk score was ineffective at predicting refractive error accurately in Africans (R^2 = 1.5%) leading to an inconsistent identification of at-risk individuals.

Overall, the results suggest that the use of genetic information to predict refractive error is a plausible option in individuals of European ancestry, and indicates a potential route to personalised myopia management. Genetic information could be evaluated to identify at-risk children who might benefit from frequent screening to identify the onset of myopia, at which point 'myopia control' treatment could be instigated. Information about a child's high genetic risk may act as an incentive for parents to bring their children for regular sight tests. In turn, the more regular screening of at-risk individuals would allow clinicians to initiate myopia management to reduce the progression of myopia at an earlier stage (which is anticipated to improve treatment efficacy (Loh et al. 2015a; Gifford et al. 2019)). Clinicians could prescribe an optical or pharmacological intervention, as described in the literature review (Section 1.2.5).

The use of genetic risk scores may also allow clinicians to suggest myopia *prevention* approaches to high risk individuals at an early age, before any other key risk factors have become established. This would likely include advising lifestyle changes such as increased time outdoors, so parents with at-risk children would be aware of the benefits

of regular time outdoors throughout childhood, not only after the child has had a significant myopic refraction change. Potentially this would reduce the incidence of myopia (Barry et al. 2016). Nevertheless, this approach may be overly cautious, since a recommendation to spend more time outdoors would be likely to benefit all children, not only those with a high genetic risk of myopia. Moreover, it would not allow the opportunity to take advantage of the stratified levels of relative myopia risk highlighted by the genetic prediction model; time outdoors would be beneficial and easy to safely implement in those with any level of increased risk of myopia. Therefore, should any other prophylactic methods for reducing myopia incidence or delaying myopia onset be found, these interventions could also be offered to parents whose children are at high genetic risk. Depending on the level of safety, and the potential benefit (due to efficacy or low cost of application), a risk-benefit analysis may be required. If it was deemed worthwhile, the new prophylactic interventions could be proposed for the individuals with the greatest risk of high myopia, as identified by the genetic risk model.

A strength of genetic risk scores for calculating myopia risk is that they can be implemented in children at any age, even before myopia has manifested i.e. before 'pre-myopia' (Flitcroft et al. 2019). This contrasts sharply with the current best available predictor of incident myopia - cycloplegic autorefraction. As future high myopes typically develop a myopic refraction before the age of 12, genetic prediction could be assessed in infancy or at birth, and would be easy to perform. However should there be any large-scale screening implemented for myopia risk (which would potentially raise ethical concerns, see below), the results should be interpreted with caution, as being categorized in a higher risk groups is not a formal diagnosis; it is only indicative of the relative chance of developing myopia for that individual compared to others. Thus, genetic prediction is useful in giving indications of relative trends amongst individuals within a defined population, but cannot exactly predict the future refractive error of any particular individual.

Furthermore, it is important to note that the accuracy obtained through genetic prediction is not as high as other methods for predicting myopia that could potentially be used as a screening protocol, such as large-scale cycloplegic autorefraction at the age of 6 to identify children with a prescription of $\leq +0.75$D (Zadnik et al. 2015). Recent evidence has shown that the inclusion of a genetic risk score does not improve the

prediction accuracy for refractive error when age, sex, and current refractive data are taken into account for Chinese children aged 6-8 years old (Chen et al. 2019). However, it may still be valuable when trying to identify at-risk children under the age of 6. It is also worth considering whether the identification of children with a higher relative risk of developing myopia may help with a targeted approach to cycloplegic autorefraction screening for children aged 6. If the genetic risk score is implemented in infancy, a more selective screening process for myopia could take place, with only genetically high risk individuals undergoing cycloplegic autorefraction at 6 years old, potentially saving many children with low myopia risk from unnecessarily undergoing this assessment, and in turn saving costs on trained eye care professionals and cycloplegic medication. It may also help identify at-risk children who would have already developed myopia by the age of 6. However, the relationship between genetic risk and refraction at 6 years old would have to be further investigated to assess the practicality of this approach and refine any parameters and thresholds that would improve its sensitivity and specificity.

This work has provided novel insights into aspects of myopia genetics; it has included working with the largest GWAS for refractive error, with the use of MTAG as an alternative form of meta-analysis to improve accuracy compared to the standard meta-analysis method. Moreover, the analyses performed in this thesis have taken advantage of larger samples of genetic data available for educational attainment, and used them alongside refractive error in a meta-analysis to show a successful improvement in genetic prediction accuracy. This has also provided evidence for how phenotypes with a relatively low genetic correlations of less than 0.35 can be combined and exploited to improve summary statistics for refractive error.

9.2 Wider Context

The theoretical limit of genetic prediction accuracy is governed by the SNP heritability estimate. For example, the most successful genetic prediction accuracy obtained for a complex trait has been for 'height', in which a genetic risk score can explain 40% of the variance in an independent sample (Lello et al. 2018). Height is highly heritable: SNP heritability is estimated to be 45 - 63% (Silventoinen et al. 2003). The SNP heritability of refractive error has been estimated to be 39% (Shah et al. 2018). Thus, genetic prediction of refractive error is much more difficult than for height.

Regarding other ophthalmic traits, studies investigating intra-ocular pressure (IOP) have developed a genetic risk score model using loci associated with IOP to predict the relative risk of developing glaucoma (through the development of IOP genetic risk scores, and applying them in a regression model for glaucoma risk). Analysing data from the UK Biobank, Khawaja et al. (2018) combined 133 loci from their GWAS meta-analysis together to create a genetic risk score, which explained up to 9% and 17% of the variation in IOP (depending on the independent validation sample used), a level of accuracy slightly greater than the genetic risk score for refractive error created in this thesis ($R^2 \approx 11\%$). The reason proposed by Khawaja et al. for the wide variation in accuracy (9% vs. 17%) in the prediction of IOP was that in the 9% accuracy sample, participants only had their IOP measured once and then this was averaged between the two eyes. In the sample with 17% accuracy, an average of 3 IOP readings per eye were averaged. Measurement error due to a single phenotypic measurement may have also been applicable to the measurement of refractive error in ALSPAC mothers, as the results obtained from the mothers was done at the end of the visit, and may have been performed only once per eye. If the ALSPAC mothers had more readings taken and averaged per eye, this may have led to an increased accuracy in genetic prediction of myopia.

Prediction of other ocular phenotypes has also been investigated with varied success. Age-related macular degeneration (AMD) genetics has been intensively studied (Arakawa et al. 2011; Fritsche et al. 2013; Seddon et al. 2013). Assuming a prevalence of 5% within the adult population, it appears that AMD has a SNP heritability of 47%, with 27% of the disease variance being explained when using known and associated genetic variants (Fritsche et al. 2016). In contrast, the trait of central corneal thickness demonstrates a poorer prediction performance ($R^2 = 8.3\%$) in individuals of European ancestry (Lu et al. 2013), despite the estimated heritability of 95% (Dimasi et al. 2010). This is likely due to the small GWAS sample size of just over 20,000 individuals used.

Other non-ocular traits that have a polygenetic aetiology show varied prediction accuracies depending on their heritability; type 2 diabetes has a heritability of approximately 25% (Almgren et al. 2011) (note that heritability varies depending on the study cohort used and on the age of onset assumed). The accuracy of genetic prediction of type 2 diabetes is no more than 10% (Voight et al. 2010). Recent genetic studies of

educational attainment have demonstrated that variants from a GWAS for self-reported educational attainment could explain 11% of the variance for years in full time education, despite the heritability of the trait being only 18% (Lee et al. 2018). This high degree of accuracy relative to the heritability of the trait may be due to the very large sample size of the GWAS for education (N = 1,100,000).

9.3 Ethical Considerations

The ability to predict phenotypes at birth would allow clinicians to screen the population early in life, before the onset of symptoms or relevant and potentially harmful physiological changes occur. Moreover, because many published GWAS reports have given insight into associated genes and variants for a variety of health conditions, a report on potential risk categories (such as high or low risk) for many of these could be supplied relatively cheaply. If desired, parents could have their children screened for many potential diseases and health conditions at any time from a simple blood test.

Although many arguments for using genetic screening stem from the proposed benefit of reducing mortality, non-life-threatening disorders such as myopia can be analysed within the same assessment. One report showed that newborn genetic screening for susceptibility to hearing loss demonstrated beneficial information to participants and clinicians in diagnosis; 13% more hearing-impaired infants were identified through genetic testing than using hearing screening alone (Wang et al. 2019). The efficiency of performing this form of assessment has also increased over time; the cost and time required for genetic screening could now be argued to be less than the time for an assessment by a healthcare professional, and has now given scope for the combination of many potential risk investigations. A recent study by Ceyhan-Birsoy et al. (2019) has shown that whole genome sequencing has been beneficial in newborns; 9.4% of the babies sampled carried a significantly higher genetic risk for childhood onset diseases identified through this method without the prospect of detection from other clinical tests or family history investigation. Thus, it could be argued that newborn screening and genetic testing should be an option provided for concerned parents and guardians, who can give informed consent to either perform genetic testing or decline the procedure. To include a risk report on the likely ocular refractive status of the child (and therefore the increased risks of secondary sight-threatening conditions they may have) alongside this test would be desired by some parents, particularly those who have had

previous family members with ocular conditions and high refractive errors. However, the ability to test genetic susceptibility during a child's lifetime, particularly at very early stages of life raises some potentially important ethical concerns that need to be considered.

Whole genome sequencing results can be difficult to interpret clinically, as in many instances the findings are inconclusive, e.g. identification of 'variants of uncertain significance' (VUS). A potentially thorny issue is the large number of potential false positives identified using genetic screening (Kingsmore 2015) i.e. if more children with greater potential risks for disease are identified, at what level of risk should clinicians intervene? Or would this screening increase unnecessary referrals and add further demand on already-strained healthcare service providers who would inherit a large body of individuals that would require closer observation? It is unlikely that all children who have a relatively greater risk of developing a condition would go on to do so, and it would be difficult to determine a threshold for said risk that would be safe in all instances of potential disease. Subsequently, increased management guidelines for these newly identified at-risk individuals may lead to theoretically excessive unnecessary interventions being performed and potentially harmful measures taken prophylactically. Making clinical judgements on the topic of what levels of risks are acceptable to be left without prophylactic intervention would be difficult to standardise, and may require further training for healthcare professionals. A risk-benefit analysis would have to be implemented for each child, and it would be important that interpretation and management using relative genetic risk be approached conservatively.

An example of a potential harmful prophylaxis process is that used in breast cancer. Genetic variants in the *BRCA1* and *BRCA2* genes increase the risk of developing breast cancer significantly: 72% and 69% of carriers for specific variants in these two genes develop breast cancer respectively, a much higher proportion than the population average (Kuchenbaecker et al. 2017). Consequently, people with a strong family history of breast cancer, or who are concerned about their risk of developing cancer have considered undergoing genetic testing to investigate whether they have inherited these predisposing genes. Should a risk mutation be identified, many clinicians advocate prophylactic mastectomies, with a significant number of people following advice and

opting to undergo this surgery (Alaofi et al. 2018). This has shown success, with 90 – 95% of individuals staying breast cancer free having undergone this preventative measure (Warner 2018). However, theory suggests that nearly a third of the women carrying risk variants would not have developed breast cancer, therefore a life-changing procedure with lasting psychosocial effects (Heidari et al. 2015) would have been performed needlessly.

Nevertheless, with respect to genetic testing for refractive error and myopia, it could be argued that the consequences of this knowledge and current interventions are minor in comparison. Myopia is not a life-threatening condition, and the interventions proposed for either avoiding myopia onset or reducing its progression are less life-changing, permanent, and easily rescindable. Although the cost of potential intervention would have to factor into any decision making (with the current use of myopia control contact lenses, spectacles, and atropine all self-funded within the UK), there is relatively little long term health/welfare concern. Contact lens wear in children poses an increased risk of eye infections, which is one of the greatest concerns. However, this risk is no higher than that for the average adult contact lens wearer, in whom contact lens use has been deemed safe (Bullimore 2017). Spending time outdoors would be cost free, and potentially have other benefits for the children's mental and physical health (Thompson Coon et al. 2011) while also having minimal risks (taking precautions for UV light from the sun for example), and thus would raise few ethical concerns.

In summary, it is likely that some ethical issues will arise from newborn genetic testing, but that investigating myopia risk would likely be one of the least controversial traits to conduct screening for. There is some evidence that offering genetic screening for concerned parents may be beneficial, ensuring to factor in what would be best for the child and parents on a case by case basis to make an informed decision. In this manner, myopia risk could be supplemented into screening platforms for a broader range of diseases, or on its own if desired.

9.4 Future Work

In this study of genetic prediction, a GWAS meta-analysis using a large combined sample of participants (mostly from UK Biobank) was used to create summary statistics that demonstrated a higher prediction efficacy for refractive error and myopia risk than previously reported. However, there is still the potential to improve the genetic

prediction of myopia further, and understand more about the genetic contribution to myopia development.

As meta-analysis using MTAG was shown to improve genetic prediction accuracy, it may be beneficial to combine the currently-used summary statistics with those from GWAS analyses of other traits. For example, combining the existing summary statistics with those from the CREAM consortium (Tedja et al. 2018) and 23&Me (Pickrell et al. 2016) would dramatically increase the effective sample size. The summary statistics for educational attainment used could also be updated to include results from a more recent GWAS of 1.1 million people to see if this would improve the genetic prediction accuracy further (Lee et al. 2018). However, this approach would likely not improve the accuracy of prediction greatly, since quantitative genetics theory suggests diminishing returns in the relationship between sample size and genetic prediction accuracy. Furthermore, although an investigation into genetic prediction of myopia in non-Europeans was done, the prediction accuracies would probably be greater if summary statistics were computed from participants with the ancestries of interest. Greater numbers of participants from other ethnicities would be useful in running separate GWAS analyses for other non-European ancestries, so that specific weights for each population - along with population-specific variants - could be identified. However, the collection of new datasets for non-European ancestry populations would be extremely costly both in time and resources.

New statistical methods could be developed that implement higher order genetic structures (such as tertiary structures of genetic material in ocular tissue such as chromatin) or utilise deep learning methods to improve prediction. Quantile regression could be used to improve accuracy for genetic prediction by adjusting the summary statistic SNP weights for individuals based on their relative genetic risk score positioning (their quantile), as a recent study has demonstrated different genetic 'effect sizes' are present for individuals with different refractive errors (currently, genetic effect sizes are assumed to be the same in every individual) (Pozarickij et al. 2019).

It may also be possible to look at other possible influences on refractive error such as epigenetics or environmental factors to give more precise prediction of refractive error (although the use of epigenetics in refractive error prediction requires validation as this has not been investigated). However, these factors are subject to change throughout

life, meaning that the prediction result would be reliable only at the time the test was conducted, and may not be as accurate if done at birth, losing its applicability in testing newborns and infants.

Furthermore, it would be interesting to investigate how genetic information and myopia predisposition can be further included in the assessment of suitability for myopia management. The following ideas may warrant investigation:

1. Exploring whether certain genetic loci are associated with a more successful outcome from interventions to reduce myopia progression.
2. Investigating if individuals with greater genetic risk obtain more or less benefit from clinical interventions (rather than those who become myopic through a greater influence from environmental factors).
3. Investigating whether individuals with specific genetic profiles will respond better to one intervention rather than another, i.e. if a precision medicine approach is warranted.

These considerations would require the initiation of clinical trials for myopia management, recruiting a large sample of participants with genetic information to detect any potential relationships.

In conclusion, a polygenic risk score has been derived that provides increased accuracy for the genetic prediction of refractive error and myopia. This polygenic risk score can be used to stratify individuals in an independent cohort based on different levels of risk, and thereby to identify those with the greatest genetic susceptibility of developing myopia. However further investigation will be required to discover if genetic susceptibility to myopia, as inferred from the genetic risk score, is predictive of treatment responses to myopia control interventions. If so, then genetic prediction may have a role in the management of childhood myopia, or in other words, a personalised medicine approach to myopia management.

References

1000 Genomes Project C, Auton A, Brooks L D, Durbin R M, Garrison E P, Kang H M, Korbel J O et al. (2015) A global reference for human genetic variation. *Nature* 526: 68-74.

Alaofi R K, Nassif M O, and Al-Hajeili M R. (2018) Prophylactic mastectomy for the prevention of breast cancer: Review of the literature. *Avicenna journal of medicine* 8: 67-77.

Aldahmesh M A, Khan A O, Alkuraya H, Adly N, Anazi S, Al-Saleh A A, Mohamed J Y et al. (2013) Mutations in LRPAP1 are associated with severe myopia in humans. *American Journal of Human Genetics* 93: 313–320.

Allen N E, Sudlow C, Peakman T, and Collins R (2014) UK biobank data: come and get it. *Science Translational Medicine* 6: 224ed.

Almgren P, Lehtovirta M, Isomaa B, Sarelin L, Taskinen M R, Lyssenko V, Tuomi T et al. (2011) Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* 54: 2811.

Andrew T, Maniatis N, Carbonaro F, Liew S H, Lau W, Spector T D, and Hammond C J. (2008) Identification and replication of three novel myopia common susceptibility gene loci on chromosome 3q26 using linkage and linkage disequilibrium mapping. *PLoS Genet*ics 4: e1000220.

Arakawa S, Takahashi A, Ashikawa K, Hosono N, Aoi T, Yasuda M, Oshima Y et al. (2011) Genome-wide association study identifies two susceptibility loci for exudative age-related macular degeneration in the Japanese population. *Nature Genetics* 43: 1001-1004.

Ashby R, Ohlendorf A, and Schaeffel F. (2009) The effect of ambient illuminance on the development of deprivation myopia in chicks. *Investigative Ophthalmology and Visual Science* 50: 5348–5354.

Au Eong K G, Tay T H, and Lim M K. (1993) Education and myopia in 110,236 young Singaporean males. *Singapore Medical Journal* 34: 489-492.

Baird P N, Schache M, and Dirani M. (2010) The Genes in Myopia (GEM) study in understanding the aetiology of refractive errors. *Progress in Retinal and Eye Research* 29: 520-542.

Balding D J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7: 781-791.

Band G, and Marchini J. (2018) BGEN: a binary file format for imputed genotype and haplotype data. *bioRxiv*: 308296.

Barry R J, Wacogne I, and Abbott J. (2016) Spending an additional 40 min outdoors each day reduces the incidence of myopia among primary school children in China. *Archives of Disease in Childhood- Education and Practice* 101: 219.

BHVI, and WHO. (2016) The report of the joint World Health Organisation - Brien Holden Vision Institure global scientific meeting on myopia. https://www.who.int/blindness/causes/MyopiaReportforWeb.pdf

Boehnke M. (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *American Journal of Human Genetics* 55: 379-390.

Bonnafous F, Fievet G, Blanchet N, Boniface M C, Carrere S, Gouzy J, Legrand L et al. (2018) Comparison of GWAS models to identify non-additive genetic control of flowering time in sunflower hybrids. *Theoretical and Applied Genetics* 131: 319-332.

Boyd A, Golding J, Macleod J, Lawlor D A, Fraser A, Henderson J, Molloy L et al. (2013) Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* 42: 111-127.

Breslow N E. (1975) Analysis of survival data under the proportional hazards model. *International Statistical Review / Revue Internationale de Statistique* 43: 45-57.

Bulik-Sullivan B, Finucane H K, Anttila V, Gusev A, Day F R, Consortium R, Genomics Consortium P et al. (2015a) An atlas of genetic correlations across human diseases and traits. *Nature Genetics* 47: 1236-1241.

Bulik-Sullivan B K, Loh P-R, Finucane H K, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, Patterson N et al. (2015b) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* 47: 291-295.

Bullimore M A. (2017) The safety of soft contact lenses in children. *Optometry and Vision Science* 94: 638-646.

Bullimore M A, and Brennan N A. (2019) Myopia control: why each diopter matters. *Optometry and Vision Science* 96: 463-465.

Bush W S, and Moore J H. (2012) Chapter 11: Genome-wide association studies. *PLOS Computational Biology* 8: e1002822.

Bycroft C, Freeman C, Petkova D, Band G, Elliott L T, Sharp K, Motyer A et al. (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562: 203-209.

Byun J, Han Y, Gorlov I P, Busam J A, Seldin M F, and Amos C I. (2017) Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure. *British Medical Council Genomics* 18: 789-789.

Calus M P L, and Vandenplas J. (2018) SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. *Genetics Selection, Evolution (GSE)* 50: 34-34.

Canela-Xandri O, Rawlik K, Woolliams J A, and Tenesa A. (2016) Improved genetic profiling of anthropometric traits using a big data approach. *PLoS One* 11: e0166755.

Cao K, Wan Y, Yusufu M, and Wang N. (2019) Significance of outdoor time for myopia prevention: a systematic review and meta-analysis based on randomized controlled trials. *Ophthalmic Research* 20: 1-9.

Carty N C, Xu J, Kurup P, Brouillette J, Goebel-Goody S M, Austin D R, Yuan P et al. (2012) The tyrosine phosphatase STEP: implications in schizophrenia and the molecular mechanism underlying antipsychotic medications. *Translational psychiatry* 2: e137.

Ceyhan-Birsoy O, Murry J B, Machini K, Lebo M S, Yu T W, Fayer S, Genetti C A et al. (2019) Interpretation of genomic sequencing results in healthy and ill newborns: results from the BabySeq project. *The American Journal of Human Genetics* 104: 76-93.

Chamberlain P, Peixoto-de-Matos S C, Logan N S, Ngo C, Jones D, and Young G. (2019) A 3-year randomized clinical trial of MiSight lenses for myopia control. *Optometry and Vision Science* 96: 556-567.

Chang C C, Chow C C, Tellier L C, Vattikuti S, Purcell S M, and Lee J J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4: 7.

Chen C Y, Scurrah K J, Stankovich J, Garoufalis P, Dirani M, Pertile K K, Richardson A J et al. (2007a) Heritability and shared environment estimates for myopia and associated ocular biometric traits: the Genes in Myopia (GEM) family study. *Human Genetics* 121: 511-520.

Chen C Y, Stankovich J, Scurrah K J, Garoufalis P, Dirani M, Pertile K K, Richardson A J et al. (2007b) Linkage replication of the MYP12 locus in common myopia. *Investigative Ophthalmology and Visual Science* 48: 4433-4439.

Chen M, Wu A, Zhang L, Wang W, Chen X, Yu X, and Wang K. (2018) The increasing prevalence of myopia and high myopia among high school students in Fenghua city, eastern China: a 15-year population-based survey. *British Medical Council Ophthalmology* 18: 159.

Chen Y, Han X, Guo X, Li Y, Lee J, and He M. (2019) Contribution of genome-wide significant single nucleotide polymorphisms in myopia prediction in children : findings from a 10-year cohort study of 1063 Chinese twin children. *Ophthalmology* 161: 33379-33387.

Chia A, Chua W H, Cheung Y B, Wong W L, Lingham A, Fong A, and Tan D. (2012) Atropine for the treatment of childhood myopia: safety and efficacy of 0.5%, 0.1%, and 0.01% doses (atropine for the treatment of myopia 2). *Ophthalmology* 119: 347-354.

Chia A, Chua W H, Wen L, Fong A, Goon Y Y, and Tan D. (2013) Atropine for the treatment of childhood myopia: changes after stopping atropine 0.01%, 0.1% and 0.5%. *American Journal of Ophthalmology* 157: 451-457.

Chia A, Lu Q S, and Tan D. (2015) Five-year clinical trial on atropine for the treatment of myopia 2: myopia control with atropine 0.01% eyedrops. *Ophthalmology* 123: 391-399.

Cho P, and Cheung S-W. (2012) Retardation of myopia in orthokeratology (ROMIO) study: a 2-Year randomized clinical trial. *Investigative Ophthalmology and Visual Science* 53: 7077-7085.

Cho P, and Cheung S W. (2017) Discontinuation of orthokeratology on eyeball elongation (DOEE). *Contact Lens and Anterior Eye* 40: 82-87.

Choi J A, Han K, Park Y M, and La T Y. (2014) Low serum 25-hydroxyvitamin D is associated with myopia in Korean adolescents. *Investigative Ophthalmology and Visual Science* 55: 2041-2047.

Chua S Y L, Sabanayagam C, Cheung Y-B, Chia A, Valenzuela R K, Tan D, Wong T-Y et al. (2016) Age of onset of myopia predicts risk of high myopia in later childhood in myopic Singapore children. *Ophthalmic and Physiological Optics* 36: 388-394.

Ciner E, Wojciechowski R, Ibay G, Bailey-Wilson J E, and Stambolian D (2008) Genomewide scan of ocular refraction in African-American families shows significant linkage to chromosome 7p15. *Genetic Epidemiology* 32: 454-463.

Collins R (2014) Making the most of UK Biobank. UK Biobank http://www.ukbiobank.ac.uk/wp-content/uploads/2014/06/0940-Collins-UKB-Frontiers-2014-1.pdf.

Cuellar-Partida G, Lu Y, Kho P F, Hewitt A W, Wichmann H E, Yazar S, Stambolian D et al. (2015) Assessing the genetic predisposition of education on myopia: a mendelian randomization study. *Genetic Epidemiology* 40: 66-72.

Cuellar-Partida G, Williams K M, Yazar S, Guggenheim J A, Hewitt A W, Williams C, Wang J J et al. (2017) Genetically low vitamin D concentrations and myopic refractive error: a Mendelian randomization study. *International Journal of Epidemiol* ogy 46: 1882-1890.

Cumberland P M, Bao Y, Hysi P G, Foster P J, Hammond C J, Rahi J S, and U. K. Biobank Eyes & Vision Consortium. (2015) Frequency and distribution of refractive error in adult life: methodology and findings of the UK Biobank study. *PLoS One* 10: e0139780.

Cumberland P M, Chianca A, and Rahi J S. (2016) Accuracy and utility of self-report of refractive error. *JAMA Ophthalmology* 134: 794-801.

Czepita M, Czepita D, and Safranow K. (2019) Role of gender in the prevalence of myopia among Polish schoolchildren. *Journal of Ophthalmology* e9748576.

Devlin B, Bacanu S-A, and Roeder K. (2004) Genomic control to the extreme. *Nature Genetics* 36: 1129-1130.

Devlin B, and Risch N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29: 311-322.

Dimasi D P, Burdon K P, and Craig J E. (2010) The genetics of central corneal thickness. *British Journal of Ophthalmology* 94: 971-976.

Djebali S, Davis C A, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A et al. (2012) Landscape of transcription in human cells. *Nature* 489: 101-108.

Dolgin E. (2015) The myopia boom. *Nature* 519: 276-278.

Dudbridge F. (2013) Power and predictive accuracy of polygenic risk scores. *PLoS Genetics* 9: e1003348.

Dudbridge F, and Gusnanto A. (2008) Estimation of significance thresholds for genome-wide association scans. *Genetic Epidemiology* 32: 227-234.

Evans D M, and Cardon L R. (2005) A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *American Journal of Human Genetics* 76: 681-687.

Fan Q, Guo X, Tideman J W, Williams K M, Yazar S, Hosseini S M, Howe L D et al. (2016a) Childhood gene-environment interactions and age-dependent effects of genetic variants associated with refractive error and myopia: The CREAM Consortium. *Scientific Reports* 6: e25853.

Fan Q, Verhoeven V J, Wojciechowski R, Barathi V A, Hysi P G, Guggenheim J A, Hohn R et al. (2016b) Meta-analysis of gene-environment-wide association scans accounting for education level identifies additional loci for refractive error. *Nature Communications* 7: e11008.

Farbrother J E, Kirov G, Owen M J, Pong-Wong R, Haley C S, and Guggenheim J A (2004) Linkage analysis of the genetic loci for high myopia on chromosomes 18p, 12q and 17q in 51 UK families. *Investigative Ophthalmology and Visual Science* 45: 2879-2885.

Flaxman S R, Bourne R R A, Resnikoff S, Ackland P, Braithwaite T, Cicinelli M V, Das A et al. (2017) Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. *The Lancet Global Health* 5: e1221-e1234.

Flister M J, Tsaih S W, O'Meara C C, Endres B, Hoffman M J, Geurts A M, Dwinell M R et al. (2013) Identifying multiple causative genes at a single GWAS locus. *Genome Research* 23: 1996-2002.

Flitcroft D I. (2012) The complex interactions of retinal, optical and environmental factors in myopia aetiology. *Progress in Retinal and Eye Research* 31: 622-660.

Flitcroft, D. I. (2014) Emmetropisation and the aetiology of refractive errors. *Eye* 28:169-179.

Flitcroft D I, He M, Jonas J B, Jong M, Naidoo K, Ohno-Matsui K, Rahi J et al. (2019) IMI – Defining and Classifying Myopia: A Proposed Set of Standards for Clinical and Epidemiologic Studies. *Investigative Ophthalmology & Visual Science* 60: M20-M30.

Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J et al. (2013) Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International J ournal of Epidemiology* 42: 97-110.

Frazer K A, Murray S S, Schork N J, and Topol E J. (2009) Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* 10: 241.

French A N, Ashby R S, Morgan I G, and Rose K A. (2013a) Time outdoors and the prevention of myopia. *Experimental Eye Research* 114: 58–68.

French A N, Morgan I G, Mitchell P, and Rose K A. (2013b) Risk factors for incident myopia in Australian schoolchildren: The Sydney Adolescent Vascular and Eye Study. *Ophthalmology* 120: 2100–2108.

Fritsche L G, Chen W, Schu M, Yaspan B L, Yu Y, Thorleifsson G, Zack D J et al. (2013) Seven new loci associated with age-related macular degeneration. *Nature Genetics* 45: 433-439.

Fritsche L G, Igl W, Bailey J N C, Grassmann F, Sengupta S, Bragg-Gresham J L, Burdon K P et al. (2016) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics* 48: 134-143.

Gabriel S B, Schaffner S F, Nguyen H, Moore J M, Roy J, Blumenstiel B, Higgins J et al. (2002) The structure of haplotype blocks in the human genome. *Science* 296: 2225-2229.

Gao X R, Huang H, Nannini D R, Fan F, and Kim H. (2018) Genome-wide association analyses identify new loci influencing intraocular pressure. *Human Molecular Genetics* 27: 2205-2213.

Ghorbani Mojarrad N, Williams C, and Guggenheim J A. (2018) A genetic risk score and number of myopic parents independently predict myopia. *Ophthalmic and Physiological Optics* 38: 492-502.

Gifford K L, Richdale K, Kang P, Aller T A, Lam C S, Liu Y M, Michaud L et al. (2019) IMI – Clinical management guidelines report. *Investigative Ophthalmology & Visual Science* 60: M184-M203.

Goh L, and Yap V B. (2009) Effects of normalization on quantitative traits in association test. *British Medical Council Bioinformatics* 10: 415.

Goldstein D B, and Weale M E. (2001) Population genomics: linkage disequilibrium holds the key. *Current Biology* 11: 576-579.

Goss D A. (2000) Nearwork and myopia. *Lancet* 356: 1456-1457.

Grosvenor T, and Scott R. (1993) 3 year changes in refraction and its components in youth-onset and early adult-onset myopia. *Optometry and Vision Science* 70: 677-683.

Guggenheim J A, Ghorbani Mojarrad N, Williams C, and Flitcroft D I. (2017) Genetic prediction of myopia: prospects and challenges. *Ophthalmic Physiological Optics* 37: 549-556.

Guggenheim J A, Hill C, and Yam T-F. (2003) Myopia, genetics and ambient lighting at night in a UK sample. *British Journal of Ophthalmology* 87: 580-582.

Guggenheim J A, Northstone K, McMahon G, Ness A R, Deere K, and Mattocks C. (2012) Time outdoors and physical activity as predictors of incident myopia in childhood: a prospective cohort study. *Invest Ophthalmology and Visual Science* 53: 2856-2865.

Guggenheim J A, St Pourcain B, McMahon G, Timpson N J, Evans D M, and Williams C. (2015) Assumption-free estimation of the genetic contribution to refractive error across childhood. *Molecular Vision* 21: 621-632.

Guggenheim J A, Williams C, Northstone K, Howe L D, Tilling K, St Pourcain B, McMahon G et al. (2014) Does vitamin D mediate the protective effects of time outdoors on myopia? Findings from a prospective birth cohort. *Investigative Ophthalmology and Visual Science* 55: 8550-8558.

Guo H, Jin X, Zhu T, Wang T, Tong P, Tian L, Peng Y et al. (2014) SLC39A5 mutations interfering with the BMP/TGF-β pathway in non-syndromic high myopia. *Journal of Medical Genetics* 51: 518-525.

Guo H, Tong P, Liu Y, Xia L, Wang T, Tian Q, Li Y et al. (2015) Mutations of P4HA2 encoding prolyl 4-hydroxylase 2 are associated with nonsyndromic high myopia. *Genetics in Medicine* 17: 300–306.

Guo X, Xiao X, Li S, Wang P, Jia X, and Zhang Q (2010) Nonsyndromic high myopia in a Chinese family mapped to MYP1: Linkage confirmation and phenotypic characterization. *Archives of Ophthalmology* 128: 1473-1479.

Hammond C J, Andrew T, Mak Y T, and Spector T D. (2004) A susceptibility locus for myopia in the normal population is linked to the PAX6 gene region on chromosome 11: A genomewide scan of dizygotic twins. *American Journal of Human Genetics* 75: 294-304.

Hartwig A, Gowen E, Charman W N, and Radhakrishnan H. (2011) Working distance and eye and head movements during near work in myopes and non-myopes. *Clinical and Experimental Optometry* 94: 536-544.

He M, Xiang F, Zeng Y, Mai J, Chen Q, Zhang J, Smith W et al. (2015) Effect of Time Spent Outdoors at School on the Development of Myopia Among Children in China: A Randomized Clinical Trial. *Journal of the American Medical Association* 314: 1142-1148.

He M, Xu M, Zhang B, Liang J, Chen P, Lee J-Y, Johnson T A et al. (2014) Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Human Molecular Genetics* 24: 1791-1800.

He M, Zeng J, Liu Y, Xu J, Pokharel G P, and Ellwein L B. (2004) Refractive error and visual impairment in urban children in southern china. *Investigative Ophthalmology and Visual Science* 45: 793-799.

Heidari M, Shahbazi S, and Ghodusi M. (2015) Evaluation of body esteem and mental health in patients with breast cancer after mastectomy. *Journal of Mid-life Health* 6: 173-177.

Hemani G, Knott S, and Haley C. (2013) An evolutionary perspective on epistasis and the missing heritability. *PLoS Genetics* 9: e1003295.

Hill W G, Goddard M E, and Visscher P M. (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* 4: e1000008.

Hirschhorn J N, and Daly M J. (2005) Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 6: 95-108.

Holden B A, Fricke T R, Wilson D A, Jong M, Naidoo K S, Sankaridurg P, Wong T Y et al. (2016) Global Prevalence of Myopia and High Myopia and Temporal Trends from 2000 through 2050. *Ophthalmology* 123: 1036-1042.

Howie B, Fuchsberger C, Stephens M, Marchini J, and Abecasis G R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* 44: 955-959.

Howie B N, Donnelly P, and Marchini J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 5: e1000529.

Howlett M H, and McFadden S A. (2006) Form-deprivation myopia in the guinea pig (Cavia porcellus). *Vision Research* 46: 267-283.

Hua W-J, Jin J-X, Wu X-Y, Yang J-W, Jiang X, Gao G-P, and Tao F-B. (2015) Elevated light levels in schools have a protective effect on myopia. *Ophthalmic and Physiological Optics* 35: 252-262.

Huang H M, Chang D S, and Wu P C. (2015) The association between near work activities and myopia in children-a systematic review and meta-analysis. *PLoS One* 10: e0140419.

Huang J, Wen D, Wang Q, McAlinden C, Flitcroft I, Chen H, Saw S M et al. (2016) Efficacy comparison of 16 interventions for myopia control in children: a network meta-analysis. *Ophthalmology* 123: 697-708.

Hui J, Peck L, and Howland H C. (1995) Correlations between familial refractive error and children's non-cycloplegic refractions. *Vision Research* 35: 1353-1358.

Hunt R, Sauna Z E, Ambudkar S V, Gottesman M M, and Kimchi-Sarfaty C. (2009) Silent (synonymous) SNPs: should we care about them? *Methods in Molecular Biology* 578: 23-39.

Hysi P G, Young T L, Mackey D A, Andrew T, Fernandez-Medarde A, Solouki A M, Hewitt A W et al. (2010) A genome-wide association study for myopia and refractive error identifies a susceptibility locus at 15q25. *Nature Genetics* 42: 902-905.

Ip J M, Saw S M, Rose K A, Morgan I G, Kifley A, Wang J J, and Mitchell P. (2008) Role of near work in myopia: findings in a sample of Australian school children. *Investigative Ophthalmology and Visual Science* 49: 2903-2910.

Iribarren R, Cortinez M F, and Chiappe J P. (2009) Age of first distance prescription and final myopic refractive error. *Ophthalmic Epidemiology* 16: 84-89.

Jiang D, Li J, Xiao X, Li S, Jia X, Sun W, Guo X et al. (2014) Detection of mutations in LRPAP1, CTSH, LEPREL1, ZNF644, SLC39A5, and SCO2 in 298 families with early-onset high myopia by exome sequencing. *Investigative Ophthalmology and Visual Science* 56: 339-345.

Jin J X, Hua W J, Jiang X, Wu X Y, Yang J W, Gao G P, Fang Y et al. (2015) Effect of outdoor activity on myopia onset and progression in school-aged children in northeast China: the Sujiatun Eye Care Study. *British Medical Council Ophthalmology* 15: 73.

Jones-Jordan L A, Sinnott L T, Manny R E, Cotter S, Kleinstein R N, Mutti D O, Twelker D et al. (2010) Early childhood refractive error and parental history of myopia as predictors of myopia. *Investigative Ophthalmology and Visual Science* 51: 115-121.

Jorde L B, and Wooding S P. (2004) Genetic variation, classification and "race". *Nature Genetics* 36: S28.

Joseph J. (1998) The Equal Environment Assumption of the Classical Twin Method: A Critical Analysis. *The Journal of Mind and Behavior* 19: 325-358.

Jung S K, Lee J H, Kakizaki H, and Jee D. (2012) Prevalence of myopia and its association with body stature and educational level in 19-year-old male conscripts in seoul, South Korea. *Investigative Ophthalmology and Visual Science* 53: 5579-5583.

Karouta C, and Ashby R S. (2015) Correlation between Light Levels and the Development of Deprivation Myopia. *Investigative Ophthalmology and Visual Science* 56: 299–309.

Keinan A, Mullikin J C, Patterson N, and Reich D. (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics* 39: 1251-1255.

Kerem B, Rommens J M, Buchanan J A, Markiewicz D, Cox T K, Chakravarti A, Buchwald M et al. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* 245: 1073-1080.

Khawaja A P, Cooke Bailey J N, Wareham N J, Scott R A, Simcoe M, Igo R P, Song Y E et al. (2018) Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. *Nature Genetics* 50: 778-782.

Khera Amit V, Chaffin M, Zekavat Seyedeh M, Collins Ryan L, Roselli C, Natarajan P, Lichtman Judith H et al. (2019) Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation* 139: 1593-1602.

Khera A V, Chaffin M, Aragam K G, Haas M E, Roselli C, Choi S H, Natarajan P et al. (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* 50: 1219-1224.

Kiefer A K, Tung J Y, Do C B, Hinds D A, Mountain J L, Francke U, and Eriksson N. (2013) Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. *PLoS Genetics* 9: e1003299.

Kim Y, Lee Y, Lee S, Kim N H, Lim J, Kim Y J, Oh J H et al. (2015) On the estimation of heritability with family-based and population-based samples. *BioMedical Research International* 2015: e671349.

Kingsmore S F. (2015) Newborn testing and screening by whole-genome sequencing. *Genetics in Medicine* 18: 214-216.

Klein A P, Duggal P, Lee K E, Klein R, Bailey-Wilson J E, and Klein B E. (2007) Confirmation of linkage to ocular refraction on chromosome 22q and identification of a novel linkage region on 1q. *Archives of Ophthalmology* 125: 80-85.

Knapp M, Wassmer G, and Baur M P (1995) The relative efficiency of the Hardy-Weinberg equilibrium-likelihood and the conditional on parental genotype-likelihood methods for candidate gene association studies. *Am J Hum Genet* 57: 1476-1485.

Korte A, and Farlow A. (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9: 29.

Krantz E M, Cruickshanks K J, Klein B E, Klein R, Huang G H, and Nieto F J. (2010) Measuring refraction in adults in epidemiological studies. *Archives of Ophthalmology* 128: 88-92.

Kuchenbaecker K B, Hopper J L, Barnes D R, Phillips K A, Mooij T M, Roos-Blom M J, Jervis S et al. (2017) Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers. *Journal of the American Medical Association* 317: 2402-2416.

Kumaran A, Htoon H M, Tan D, and Chia A. (2015) Analysis of changes in refraction and biometry of atropine- and placebo-treated eyes. *Investigative Ophthalmology and Visual Science* 56: 5650-5655.

Kurup P K, Xu J, Videira R A, Ononenyi C, Baltazar G, Lombroso P J, and Nairn A C. (2015) STEP61 is a substrate of the E3 ligase parkin and is upregulated in Parkinson's disease. *Proceedings of the National Academy of Sciences of the United States of America* 112: 1202-1207.

Ladstätter S, and Tachibana-Konwalski K. (2016) A surveillance mechanism ensures repair of dna lesions during zygotic reprogramming. *Cell* 167: 1774-1787.

LaFramboise T. (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research* 37: 4181-4193.

Lam C Y, Tam P O, Fan D S, Fan B J, Wang D Y, Lee C W, Pang C P et al. (2008) A genome-wide scan maps a novel high myopia locus to 5p15. *Investigative Ophthalmology and Visual Science* 49: 3768-3778.

Lam, C. S. Y. et al. (2019) Defocus Incorporated Multiple Segments (DIMS) spectacle lenses slow myopia progression: a 2-year randomised clinical trial. *British Journal of Ophthalmol*, 2018-313739.

Lander E S, and Schork N J. (1994) Genetic dissection of complex traits. *Science* 265: 2037-2048.

Lasker G. (1956) Anthropological and ophthalmological studies on the Angmagssalik Eskimos. *Human Biology* 28: 385-387.

Lee I, Blom U M, Wang P I, Shim J E, and Marcotte E M. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research* 21: 1109-1121.

Lee J J, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, Nguyen-Viet T A et al. (2018) Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics* 50: 1112-1121.

Lello L, Avery S G, Tellier L, Vazquez A I, de los Campos G, and Hsu S D H. (2018) Accurate Genomic Prediction of Human Height. *Genetics* 210: 477.

Li J, Jiang D, Xiao X, Li S, Jia X, Sun W, Guo X et al. (2015) Evaluation of 12 myopia-associated genes in Chinese patients with high myopia. *Investigative Ophthalmology and Visual Science* 56: 722-729.

Li Y J, Goh L, Khor C C, Fan Q, Yu M, Han S, Sim X et al. (2011) Genome-wide association studies reveal genetic variants in CTNND2 for high myopia in Singapore Chinese. *Ophthalmology* 118: 368–375.

Lim L T, Gong Y, Ah-Kee E Y, Xiao G, Zhang X, and Yu S. (2014) Impact of parental history of myopia on the development of myopia in mainland china school-aged children. *Ophthalmology and Eye Diseases* 6: 31-35.

Lin L L, Shih Y F, Hsiao C K, and Chen C J. (2004) Prevalence of myopia in Taiwanese schoolchildren: 1983 to 2000. *Annals Academy of Medicine Singapore* 33: 27-33.

Lin Z, Vasudevan B, Ciuffreda K J, Wang N L, Zhang Y C, Rong S S, Qiao L Y et al. (2013) Nearwork-induced transient myopia and parental refractive error. *Optometry and Vision Science* 90: 507-516.

Linck E, and Battey C J. (2019) Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources* 19: 639-647.

Listgarten J, Lippert C, Kadie C M, Davidson R I, Eskin E, and Heckerman D. (2012) Improved linear mixed models for genome-wide association studies. *Nature Methods* 9: 525-526.

Liu J, Wang K, Ma S, and Huang J. (2013) Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method. *Statistics and its Interface* 6: 99-115.

Lobo I. (2008) Environmental influences on gene expression. *Nature Education: Nature*. p39

Lodish H B A, Matsudaira P, Kaiser C, Krieger M, Scott M, Zipursky L, Darnell J. (2004) *Molecular Cell Biology*. 5th Edition ed. San Francisco: W.H.Freeman.

Logan N S, Shah P, Rudnicka A R, Gilmartin B, and Owen C G. (2011) Childhood ethnic differences in ametropia and ocular biometry: the Aston Eye Study. *Ophthalmic and Physiological Optics* 31: 550–558.

Loh K L, Lu Q, Tan D, and Chia A. (2015a) Risk Factors for Progressive Myopia in Atropine Therapy for Myopia (ATOM 1 Study). *American Journal of Ophthalmology* 159: 945-949.

Loh P-R, Kichaev G, Gazal S, Schoech A P, and Price A L. (2018) Mixed-model association for biobank-scale datasets. *Nature Genetics* 50: 906-908.

Loh P-R, Tucker G, Bulik-Sullivan B K, Vilhjalmsson B J, Finucane H K, Salem R M, Chasman D I et al. (2015b) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* 47: 284-290.

Lombroso P J, Murdoch G, and Lerner M. (1991) Molecular characterization of a protein-tyrosine-phosphatase enriched in striatum. *Proceedings of the National Academy of Sciences of the United States of America* 88: 7242-7246.

Loughman J, and Flitcroft D I. (2016) The acceptability and visual impact of 0.01% atropine in a Caucasian population. *British Journal of Ophthalmology* 100: 1525-1529.

Lu Y, Vitart V, Burdon K P, Khor C C, Bykhovskaya Y, Mirshahi A, Hewitt A W et al. (2013) Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nature Genetics* 45: 155-163.

Lyhne N, Sjolie A K, Kyvik K O, and Green A. (2001) The importance of genes and environment for ocular refraction and its determiners: a population based study among 20-45 year old twins. *British Journal of Ophthalmology* 85: 1470-1476.

Ma J H, Shen S H, Zhang G W, Zhao D S, Xu C, Pan C M, Jiang H et al. (2010) Identification of a locus for autosomal dominant high myopia on chromosome 5p13.3-p15.1 in a Chinese family. *Molecular Vision* 16: 2043-2054.

Marcus M W, de Vries M M, Montolio F G, and Jansonius N M. (2011) Myopia as a risk factor for open-angle glaucoma: A systematic review and meta-analysis. *Ophthalmology* 118: 1989-1994.

Marees A T, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, and Derks E M (2018) A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research* 27: e1608.

Marquez-Luna C, Loh P R, and Price A L. (2017) Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology* 41: 811-823.

McCarroll S A, and Altshuler D M. (2007) Copy-number variation and association studies of human disease. *Nature Genetics* 39: 37-42.

McCarthy C S, Megaw P, Devadas M, and Morgan I G (2007) Dopaminergic agents affect the ability of brief periods of normal vision to prevent form-deprivation myopia. *Experimental Eye Research* 84: 100-107.

McClellan J, and King M-C. (2010) Genetic heterogeneity in human disease. *Cell* 141: 210-217.

Mersha T B, and Abebe T. (2015) Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. *Human Genomics* 9: 1.

Millodot M. (2014) *Dictionary of Optometry and Visual Science*. Elsevier Health Sciences UK.

Mimouni M, Zoller L, Horowitz J, Wygnanski-Jaffe T, Morad Y, and Mezer E. (2016) Cycloplegic autorefraction in young adults: is it mandatory? *Graefes Archive of Clinical and Experimental Ophthalmology* 254: 395-398.

Monir M M, and Zhu J. (2017) Comparing gwas results of complex traits using full genetic model and additive models for revealing genetic architecture. *Scientific Reports* 7: e38600.

Morgan I, and Rose K. (2005) How genetic is school myopia? *Progress in Retinal and Eye Research* 24: 1-38.

Morgan I G, French A N, Ashby R S, Guo X, Ding X, He M, and Rose K A. (2017) The epidemics of myopia: Aetiology and prevention. *Progress in Retinal Eye Research* 62: 134-149.

Morgan I G, Ohno-Matsui K, and Saw S M. (2012) Myopia. *Lancet* 379: 1739-1748.

Morgan R W, Speakman J S, and Grimshaw S E. (1975) Inuit myopia: an environmentally induced "epidemic"? *Canadian Medical Association journal* 112: 575-577.

Mountjoy E, Davies N M, Plotnikov D, Smith G D, Rodriguez S, Williams C E, Guggenheim J A et al. (2018) Education and myopia: assessing the direction of causality by mendelian randomisation. *British Medical Journal* 6: 361.

Mullaney J M, Mills R E, Pittard W S, and Devine S E. (2010) Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics* 19: 131-136.

Munafò, M. R. et al. (2018) Collider scope: when selection bias can substantially influence observed associations. *International Jour of Epidemiology* 47: 226-235.

Mutti D O, Hayes J R, Mitchell G L, Jones L A, Moeschberger M L, Cotter S A, Kleinstein R N et al. (2007) Refractive error, axial length, and relative peripheral refractive error before and after the onset of myopia. *Investigative Ophthalmology & Visual Science* 48: 2510-2519.

Mutti D O, and Marks A R. (2011) Blood levels of Vitamin D in teens and young adults with myopia. *Optometry and Vision Science* 88: 377-382.

Mutti D O, Mitchell G L, Jones L A, Friedman N E, Frane S L, Lin W K, Moeschberger M L et al. (2005) Axial growth and changes in lenticular and corneal power during emmetropization in infants. *Investigative Ophthalmology & Visual Science* 46: 3074-3080.

Mutti D O, Mitchell G L, Moeschberger M L, Jones L A, and Zadnik K (2002) Parental myopia, near work, school achievement, and children's refractive error. *Investigative Ophthalmology and Visual Science* 43: 3633-3640.

Mutti D O, and Zadnik K. (2000) Age-related decreases in the prevalence of myopia: Longitudinal change or cohort effect? *Investigative Ophthalmology and Visual Science* 41: 2103-2107.

Mutti D O, and Zadnik K. (2009) Has near work's star fallen? *Optometry and Vision Science* 86: 76-78.

Mutti D O, Zadnik K, and Murphy C J. (1998) The effect of continuous light on refractive error and the ocular components of the rat. *Experimental Eye Research* 67: 631-636.

Nakanishi H, Yamada R, Gotoh N, Hayashi H, Yamashiro K, Shimada N, Ohno-Matsui K et al. (2009) A Genome-Wide Association Analysis Identified a Novel Susceptible Locus for Pathological Myopia at 11q24.1. *PLoS Genet* 5: e1000660.

Nallasamy S, Paluru P C, Devoto M, Wasserman N F, Zhou J, and Young T L (2007) Genetic linkage study of high-grade myopia in a Hutterite population from South Dakota. *Mol Vis* 13: 229-236.

Ngo C S, Pan C W, Finkelstein E A, Lee C F, Wong I B, Ong J, Ang M et al. (2014) A cluster randomised controlled trial evaluating an incentive-based outdoor physical activity programme to increase outdoor time and prevent myopia in children. *Ophthalmic Physiol Opt* 34: 362-368.

Nickla D L, and Totonelly K. (2016) Brief light exposure at night disrupts the circadian rhythms in eye growth and choroidal thickness in chicks. *Experimental Eye Research* 146: 189-195.

Northstone K, Guggenheim J A, Howe L D, Tilling K, Paternoster L, Kemp J P, McMahon G et al. (2013) Body stature growth trajectory during childhood and the development of myopia. *Ophthalmology* 120: 1064-1073.

Norton T T, and Siegwart J T, Jr. (2013) Light levels, refractive development, and myopia - A speculative review. *Experimental Eye Research* 114: 48–57.

Nurnberg G, Jacobi F K, Broghammer M, Becker C, Blin N, Nurnberg P, Stephani U et al. (2008) Refinement of the MYP3 locus on human chromosome 12 in a German family with Mendelian autosomal dominant high-grade myopia by SNP array mapping. *International Journal of Molecular Medicine* 21: 429-438.

O'Donoghue L, McClelland J F, Logan N S, Rudnicka A R, Owen C G, and Saunders K J. (2010) Refractive error and visual impairment in school children in Northern Ireland. *British Journal of Ophthalmology* 94: 1155-1159.

Okbay A, Beauchamp J P, Fontana M A, Lee J J, Pers T H, Rietveld C A, Turley P et al. (2016) Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 533: 539-542.

Orazio J, Jarrett S, Amaro-Ortiz A, and Scott T. (2013) UV Radiation and the Skin. *International Journal of Molecular Sciences* 14: 12222-12248.

Paget S, Julia S, Vitezica Z G, Soler V, Malecaze F, and Calvas P. (2008) Linkage analysis of high myopia susceptibility locus in 26 families. *Molecular Vision* 14: 2566-2574.

Paluru P, Ronan S M, Heon E, Devoto M, Wildenberg S C, Scavello G, Holleschau A et al. (2003) New locus for autosomal dominant high myopia maps to the long arm of chromosome 17. *Investigative Ophthalmology and Visual Science* 44: 1830-1836.

Paluru P C, Nallasamy S, Devoto M, Rappaport E F, and Young T L. (2005) Identification of a novel locus on 2q for autosomal dominant high-grade myopia. *Investigative Ophthalmology and Visual Science* 46: 2300-2307.

Pan C-W, Ramamurthy D, and Saw S-M. (2012) Worldwide prevalence and risk factors for myopia. *Ophthalmic and Physiological Optics* 32: 3-16.

Pan C W, Dirani M, Cheng C Y, Wong T Y, and Saw S M. (2015) The age-specific prevalence of myopia in Asia: a meta-analysis. *Optometry and Visual Science* 92: 258-266.

Pan C W, Qian D J, and Saw S M. (2017) Time outdoors, blood vitamin D status and myopia: a review. *Photochemical and Photobiological Science* 16: 426-432.

Pärssinen O, Kauppinen M, and Viljanen A. (2014) The progression of myopia from its onset at age 8–12 to adulthood and the influence of heredity and external factors on myopic progression. A 23-year follow-up study. *Acta Ophthalmolgie* 92: 730-739.

Patterson N, Price A L, and Reich D. (2006) Population structure and eigenanalysis. *PLoS Genet* 2: e190.

Peet J A, Cotch M F, Wojciechowski R, Bailey-Wilson J E, and Stambolian D. (2007) Heritability and familial aggregation of refractive error in the Old Order Amish. *Investigative Ophthalmology and Visual Science* 48: 4002-4006.

Perry D J, Wasserfall C H, Oram R A, Williams M D, Posgai A, Muir A B, Haller M J et al. (2018) Application of a genetic risk score to racially diverse type 1 diabetes populations demonstrates the need for diversity in risk-modeling. *Scientific Reports* 8: 4529.

Pickrell J K, Berisa T, Liu J Z, Segurel L, Tung J Y, and Hinds D A. (2016) Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics* 48: 709-717.

Plotnikov D, Shah R L, Rodrigues J N, Cumberland P M, Rahi J S, Hysi P G, Atan D et al. (2019) A commonly occurring genetic variant within the NPLOC4–TSPAN10–PDE6G gene cluster is associated with the risk of strabismus. *Human Genetics* 138: 723-737.

Polling J R, Kok R G, Tideman J W, Meskat B, and Klaver C C. (2016) Effectiveness study of atropine for progressive myopia in Europeans. *Eye* 30: 998-1004.

Pozarickij A, Williams C, Hysi P G, Guggenheim J A, Aslam T, Barman S A, Barrett J H et al. (2019) Quantile regression analysis reveals widespread evidence for gene-environment or gene-gene interactions in myopia development. *Communications Biology* 2: 167.

Price A L, Patterson N J, Plenge R M, Weinblatt M E, Shadick N A, and Reich D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38: 904-909.

Pritchard J K. (2001) Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics* 69: 124-137.

Purcell S. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81: 559-575.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M A R, Bender D, Maller J et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81: 559-575.

Quinn G E, Shin C H, Maguire M G, and Stone R A. (1999) Myopia and ambient lighting at night. *Nature* 399: 113-114.

Rabbee N, and Speed T P. (2006) A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* 22: 7-12.

Ratnamala U, Lyle R, Raval R, Singh R, Vishnupriya S, Himabindu P, Rao V V et al. (2011) Refinement of the X-linked nonsyndromic high-grade myopia locus (MYP1) on Xq28 and exclusion of thirteen known positional candidate genes by direct sequencing. *Investigative Ophthalmology and Visual Science* 52: 6814-6819.

Rawlik K, Canela-Xandri O, and Tenesa A. (2016) Evidence for sex-specific genetic architectures across a spectrum of human complex traits. *Genome Biology* 17: 166.

Reich D E, and Goldstein D B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology* 20: 4-16.

Richardson K, and Norgate S. (2005) The equal environments assumption of classical twin studies may not hold. *British Journal of Educational Psychology* 75: 339-350.

Richmond R C, Timpson N J, Felix J F, Palmer T, Gaillard R, McMahon G, Davey Smith G et al. (2017) Using genetic variation to explore the causal effect of maternal pregnancy adiposity on future offspring adiposity: a mendelian randomisation study. *PLoS Medicine* 14: e1002221.

Risch N. (1990) Linkage strategies for genetically complex traits & multilocus models. *American Journal of Human Genetics* 46: 222-228.

Risch N, and Merikangas K. (1996) The future of genetic studies of complex human diseases. *Science* 273: 1516-1517.

Risch N, and Teng J. (1997) Design and analysis of linkage disequilibrium studies for complex human diseases. *American Journal of Human Genetics* 61: 1707.

Rose K A, Morgan I G, Ip J, Kifley A, Huynh S, Smith W, and Mitchell P. (2008) Outdoor activity reduces the prevalence of myopia in children. *Ophthalmology* 115: 1279-1285.

Rosenfield M, and Gilmartin B. (1998) Myopia and nearwork: Causation or merely association? In: Rosenfield, M, and Gilmartin, B [eds.] *Myopia and nearwork*. Oxford, UK: Butterworth-Heinemann.

Rosner M, and Belkin M. (1987) Intelligence, education and myopia in males. *Archives of Ophthalmology* 105: 1508-1511.

Rudnicka A R, Kapetanakis V V, Wathern A K, Logan N S, Gilmartin B, Whincup P H, Cook D G et al. (2016) Global variations and time trends in the prevalence of childhood myopia, a systematic review and quantitative meta-analysis: implications for aetiology and early prevention. *British Journal of Ophthalmology* 100: 882-890.

Rydzanicz M, Nath S K, Sun C, Podfigurna-Musielak M, Frajdenberg A, Mrugacz M, Winters D et al. (2011) Identification of novel suggestive loci for high-grade myopia in Polish families. *Molecular Vision* 17: 2028-2039.

Sanfilippo P G, Chu B S, Bigault O, Kearns L S, Boon M Y, Young T L, Hammond C J et al. (2014) What is the appropriate age cut-off for cycloplegia in refraction? *Acta Ophthalmologie* 92: 458-462.

Sanfilippo P G, Hewitt A W, Hammond C J, and Mackey D A. (2010) The heritability of ocular traits. *Survey of Ophthalmology* 55: 561-583.

Sanfilippo P G, Medland S E, Hewitt A W, Kearns L S, Ruddle J B, Sun C, Hammond C J et al. (2011) Ophthalmic phenotypes and the representativeness of twin data for the general population. *Investigative Ophthalmology and Visual Science* 52: 5565-5572.

Sankaridurg P, Bakaraju R C, Naduvilath T, Chen X, Weng R, Tilia D, Xu P et al. (2019) Myopia control with novel central and peripheral plus contact lenses and extended depth of focus contact lenses: 2 year results from a randomised clinical trial. *Ophthalmic and Physiological Optics* 39: 294-307.

Sankaridurg P, He X, Naduvilath T, Lv M, Ho A, Smith E, 3rd, Erickson P et al. (2017) Comparison of noncycloplegic and cycloplegic autorefraction in categorizing refractive error data in children. *Acta Ophthalmologie* 95: 633-640.

Saw S-M, Chua W-H, Hong C-Y, Wu H-M, Chan W-Y, Chia K-S, Stone R A et al. (2002) Nearwork in early-onset myopia. *Investigative Ophthalmology and Visual Science* 43: 332-339.

Saw S M, Shankar A, Tan S B, Taylor H, Tan D T, Stone R A, and Wong T Y. (2006) A cohort study of incident myopia in Singaporean children. *Investigative Ophthalmology and Visual Science* 47: 1839-1844.

Saw S M, Tan S B, Fung D, Chia K S, Koh D, Tan D T, and Stone R A. (2004) IQ and the association with myopia in children. *Investigative Ophthalmology and Visual Science* 45: 2943-2948.

Schache M, and Baird P N. (2013) The Australian Twin Registry as a resource for genetic studies into ophthalmic traits. *Twin Research and Human Genetics* 16: 52-57.

Schache M, Chen C Y, Pertile K K, Richardson A J, Dirani M, Mitchell P, and Baird P N. (2009) Fine mapping linkage analysis identifies a novel susceptibility locus for myopia on chromosome 2q37 adjacent to but not overlapping MYP12. *Molecular Vision* 15: 722-730.

Seddon J M, Yu Y, Miller E C, Reynolds R, Tan P L, Gowrisankar S, Goldstein J I et al. (2013) Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. *Nature Genetics* 45: 1366-1370.

Shah R L, Guggenheim J A, Eye U K B, and Vision C. (2018) Genome-wide association studies for corneal and refractive astigmatism in UK Biobank demonstrate a shared role for myopia susceptibility loci. *Human Genetics* 137: 881-896.

Shah R L, Huang Y, Guggenheim J A, and Williams C. (2017) Time outdoors at specific ages during early childhood and the risk of incident myopia. *Investigative Ophthalmology & Visual Science* 58: 1158-1166.

Sharp A J, Locke D P, McGrath S D, Cheng Z, Bailey J A, Vallente R U, Pertz L M et al. (2005) Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics* 77: 78-88.

Shi Y, Li Y, Zhang D, Zhang H, Lu F, Liu X, He F et al. (2011a) Exome Sequencing Identifies ZNF644 Mutations in High Myopia. *PLoS Genetics* 7: e1002084.

Shi Y, Qu J, Zhang D, Zhao P, Zhang Q, Tam P O, Sun L et al. (2011b) Genetic variants at 13q12.12 are associated with high myopia in the han chinese population. *American Journal of Human Genetics* 88: 805-813.

Silventoinen K, Sammalisto S, Perola M, Boomsma D I, Cornes B K, Davis C, Dunkel L et al. (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Research* 6: 399-408.

Smith E L, Hung L-F, and Huang J (2012) Protective effects of high ambient lighting on the development of form-deprivation myopia in rhesus monkeys. *Investigative Ophthalmology and Visual Science* 53: 421-428.

Solouki A M, Verhoeven V J, van Duijn C M, Verkerk A J, Ikram M K, Hysi P G, Despriet D D et al. (2010) A genome-wide association study identifies a susceptibility locus for refractive errors and myopia at 15q14. *Nature Genetics* 42: 897-901.

Stambolian D, Ciner E B, Reider L C, Moy C, Dana D, Owens R, Schlifka M et al. (2005) Genome-wide scan for myopia in the Old Order Amish. *American Journal of Ophthalmology* 5: 469-476.

Stambolian D, Ibay G, Reider L, Dana D, Moy C, Schlifka M, Holmes T et al. (2004) Genomewide linkage scan for myopia susceptibility loci among Ashkenazi Jewish families shows evidence of linkage on chromosome 22q12. *American Journal of Human Genetics* 75: 448-459.

Stone R A, Pardue M T, Iuvone P M, and Khurana T S (2013) Pharmacology of myopia and potential role for intrinsic retinal circadian rhythms. *Experimental Eye Research* 114: 35–47.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P et al. (2015) UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* 12: e1001779.

Tang W C, Yap M K, and Yip S P. (2008) A review of current approaches to identifying human genes involved in myopia. *Clinical and Experimental Optometry* 91: 4-22.

Tateno Y, Komiyama T, Katoh T, Munkhbat B, Oka A, Haida Y, Kobayashi H et al. (2014) Divergence of east asians and europeans estimated using male- and female-specific genetic markers. *Genome Biology and Evolution* 6: 466-473.

Taylor M, Simpkin A J, Haycock P C, Dudbridge F, and Zuccolo L. (2016) Exploration of a polygenic risk score for alcohol consumption: a longitudinal analysis from the alspac cohort. *PLoS One* 11: e0167360.

Tedja M S, Wojciechowski R, Hysi P G, Eriksson N, Furlotte N A, Verhoeven V J M, Iglesias A I et al. (2018) Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. *Nature Genetics* 50: 834-848.

Tenesa A, and Haley C S. (2013) The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics* 14: 139-149.

Terwilliger J D, and Weiss K M. (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* 9: 578-594.

The Genomes Project C, Auton A, Abecasis G R, Altshuler D M, Durbin R M, Abecasis G R, Bentley D R et al. (2015) A global reference for human genetic variation. *Nature* 526: 68-74.

The International HapMap C, Altshuler D M, Gibbs R A, Peltonen L, Altshuler D M, Gibbs R A, Peltonen L et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52-58.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* 437: 1299-1320.

The International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.

Thompson Coon J, Boddy K, Stein K, Whear R, Barton J, and Depledge M H. (2011) Does participating in physical activity in outdoor natural environments have a greater effect on physical and mental wellbeing than physical activity indoors? A systematic review. *Environment Science and Technology* 45: 1761-1772.

Thomson R, and McWhirter R. (2017) Adjusting for familial relatedness in the analysis of gwas data. *Methods in Molecular Biology* 1526: 175-190.

Tideman J W L, Polling J R, Voortman T, Jaddoe V W V, Uitterlinden A G, Hofman A, Vingerling J R et al. (2016) Low serum vitamin D is associated with axial length and risk of myopia in young children. *European Journal of Epidemiology* 31: 491-499.

Torkamani A, Wineinger N E, and Topol E J. (2018) The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* 19: 581-590.

Tran-Viet K N, St Germain E, Soler V, Powell C, Lim S H, Klemm T, Saw S M et al. (2012) Study of a US cohort supports the role of ZNF644 and high-grade myopia susceptibility. *Molecular Vision* 18: 937-944.

Trehearne A. (2016) Genetics, lifestyle and environment. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz* 59: 361-367.

Turley P, Walters R K, Maghzian O, Okbay A, Lee J J, Fontana M A, Nguyen-Viet T A et al. (2018) Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics* 50: 229-237.

Ueda E, Yasuda M, Fujiwara K, Hashimoto S, Ohno-Matsui K, Hata J, Ishibashi T et al. (2019) Trends in the prevalence of myopia and myopic maculopathy in a japanese population: the hisayama study. *Investigative Ophthalmology and Visual Science* 60: 2781-2786.

UK Biobank Team (2007) *UK Biobank: Protocol for a large-scale prospective epidemiological resource.* http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf Accessed: May 1, 2018.

UK Biobank Team (2018) *Biobank Data-Field Questionnaire Responses - Optical Questions.* http://biobank.ctsu.ox.ac.uk/crystal/search.cgi?wot=0&srch=glasses&sta0=on&sta1=on&sta2=on&sta3=on&str0=on&str3=on&fit0=on&fvt11=on&fvt21=on&fvt22=on&fvt31=on&fvt41=on&fvt51=on&fvt61=on&fvt101=on Accessed: May 2, 2018

Veerappan S, Schache M, Pertile K K, Islam F M, Chen C Y, Mitchell P, Dirani M et al. (2009) The retinoic acid receptor alpha (RARA) gene is not associated with myopia, hypermetropia, and ocular biometric measures. *Molecular Vision* 15: 1390-1397.

Verhoeven V J M, Hysi P G, Wojciechowski R, Fan Q, Guggenheim J A, Hohn R, MacGregor S et al. (2013) Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nature Genetics* 45: 314-318.

Verhoeven V J M, Hysi P G, Wojciechowski R, Fan Q, Kiefer A K, Klaver C C W, Hammond C J et al. (2014) Genome-wide mega-analysis on myopia and refractive error in CREAM and 23andMe. *ARVO Meeting Abstracts* 55: 839.

Vilhjálmsson Bjarni J, Yang J, Finucane Hilary K, Gusev A, Lindström S, Ripke S, Genovese G et al. (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics* 97: 576-592.

Visscher P M, Andrew T, and Nyholt D R. (2008a) Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. *European Journal of Human Genetics* 16: 387-390.

Visscher P M, Brown M A, McCarthy M I, and Yang J (2012a) Five years of GWAS discovery. *American Journal of Human Genetics* 90: 7-24.

Visscher P M, Hill W G, and Wray N R (2008b) Heritability in the genomics era--concepts and misconceptions. *Nature Reviews Genetics* 9: 255-266.

Visscher P M, Medland S E, Ferreira M A, Morley K I, Zhu G, Cornes B K, Montgomery G W et al. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics* 2: e41.

Visscher P M, Yang J, and Goddard M E. (2012b) A commentary on 'common snps explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Research and Human Genetics* 13: 517-524.

Vitale S, Sperduto R D, and Ferris F L. (2009) Increased Prevalence of Myopia in the United States Between 1971-1972 and 1999-2004. *Archives of Ophthalmology* 127: 1632-1639.

Voight B F, Scott L J, Steinthorsdottir V, Morris A P, Dina C, Welch R P, Zeggini E et al. (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* 42: 579-589.

Wall J D. (2017) Inferring human demographic histories of non-african populations from patterns of allele sharing. *American Journal of Human Genetics* 100: 766-772.

Wallman, J. (1993) Retinal control of eye growth and refraction. *Progress in Retinal Research* 12:133-153.

Walsh S, Lindenbergh A, Zuniga S B, Sijen T, de Knijff P, Kayser M, and Ballantyne K N. (2011) Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. *Forensic Science International Genetics* 5: 464-471.

Wang D G, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280: 1077-1082.

Wang Q, Xiang J, Sun J, Yang Y, Guan J, Wang D, Song C et al. (2019) Nationwide population genetic screening improves outcomes of newborn screening for hearing loss in China. *Genetics in Medicine* 21: 2231-2238.

Warner E. (2018) Screening BRCA1 and BRCA2 mutation carriers for breast cancer. *Cancers* 10: 477.

Weiss K M, and Clark A G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics* 18: 19-24.

Wen G, Tarczy-Hornoch K, McKean-Cowdin R, Cotter S A, Borchert M, Lin J, Kim J et al. (2013) Prevalence of myopia, hyperopia, and astigmatism in non-Hispanic white and Asian children: multi-ethnic pediatric eye disease study. *Ophthalmology* 120: 2109-2116.

Wild S, Roglic G, Green A, Sicree R, and King H. (2004) Global prevalence of diabetes. *Diabetes Care* 27: 1047-1053.

Wiliams A L, Patterson N, Glessner J, Hakonarson H, and Reich D. (2012) Phasing of many thousands of genotyped samples. *American Journal of Human Genetics* 91: 238-251.

Willer C J, Li Y, and Abecasis G R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26: 2190-2191.

Williams C, Miller L, Northstone K, and Sparrow J M (2008) The use of non-cycloplegic autorefraction data in general studies of children's development. *British Journal of Ophthalmology* 92: 723-724.

Williams K M, Bertelsen G, Cumberland P, Wolfram C, Verhoeven V J M, Anastasopoulos E, Buitendijk G H S et al. (2015a) Increasing prevalence of myopia in europe and the impact of education. *Ophthalmology* 122: 1489-1497.

Williams K M, Hysi P G, Nag A, Yonova-Doing E, Venturini C, and Hammond C J. (2013) Age of myopia onset in a British population-based twin cohort. *Ophthalmic and Physiological Optics* 33: 339-345.

Williams K M, Verhoeven V J, Cumberland P, Bertelsen G, Wolfram C, Buitendijk G H, Hofman A et al. (2015b) Prevalence of refractive error in Europe: the European Eye Epidemiology (E3) Consortium. *European Journal of Epidemiology* 30: 305–315.

Witkovsky P. (2004) Dopamine and retinal function. *Documenta Ophthalmologia* 108: 17-40.

Wojciechowski R. (2011) Nature and nurture: the complex genetics of myopia and refractive error. *Clinical Genetics* 79: 301-320.

Wojciechowski R, Congdon N, Bowie H, Munoz B, Gilbert D, and West S K. (2005) Heritability of refractive error and familial aggregation of myopia in an elderly american population. *Investigative Ophthalmology and Visual Science* 46: 1588-1592.

Wojciechowski R, Moy C, Ciner E, Ibay G, Reider L, Bailey-Wilson J E, and Stambolian D. (2006) Genomewide scan in Ashkenazi Jewish families demonstrates evidence of linkage of ocular refraction to a QTL on chromosome 1p36. *Human Genetics* 119: 389-399.

Wolffsohn J S, Flitcroft D I, Gifford K L, Jong M, Jones L, Klaver C C W, Logan N S et al. (2019) IMI – Myopia Control Reports Overview and IntroductionIMI – Myopia Control Reports Overview and Introduction. *Investigative Ophthalmology & Visual Science* 60: M1-M19.

Wong T Y, Ferreira A, Hughes R, Carter G, and Mitchell P. (2014) Epidemiology and disease burden of pathologic myopia and myopic choroidal neovascularization: an evidence-based systematic review. *American Journal of Ophthalmology* 157: 9-25.

Wu H M, Seet B, Yap E P H, Saw S M, Lim T H, and Chia K S. (2001) Does education explain ethnic differences in myopia prevalence? A population-based study of young adult males in Singapore. *Optometry and Vision Science* 78: 234-239.

Wu L J, You Q S, Duan J L, Luo Y X, Liu L J, Li X, Gao Q et al. (2015) Prevalence and associated factors of myopia in high-school students in Beijing. *PLoS One* 10: e0120764.

Wu M M, and Edwards M H. (1999) The effect of having myopic parents: an analysis of myopia in three generations. *Optometry and Vision Science* 76: 387-392.

Wu P C, Chen C T, Lin K K, Sun C C, Kuo C N, Huang H M, Poon Y C et al. (2018) Myopia prevention and outdoor light intensity in a school-based cluster randomized trial. *Ophthalmology* 125: 1239-1250.

Wu P C, Tsai C L, Wu H L, Yang Y H, and Kuo H K. (2013) Outdoor activity during class recess reduces myopia onset and progression in school children. *Ophthalmology* 120: 1080-1085.

Xiang X, Wang T, Tong P, Li Y, Guo H, Wan A, Xia L et al. (2014) New ZNF644 mutations identified in patients with high myopia. *Molecular Vision* 20: 939-946.

Xiao X, Li S, Jia X, Guo X, and Zhang Q. (2016) X-linked heterozygous mutations in ARR3 cause female-limited early onset high myopia. *Molecular Vision* 22: 1257-1266.

Xiong S, Sankaridurg P, Naduvilath T, Zang J, Zou H, Zhu J, Lv M et al. (2017) Time spent in outdoor activities in relation to myopia prevention and control: a meta-analysis and systematic review. *Acta Ophthalmolgica* 95: 551-566.

Yang C-H, Huang C-C, and Hsu K-S. (2012) A Critical role for protein tyrosine phosphatase nonreceptor type 5 in determining individual susceptibility to develop stress-related cognitive and morphological changes. *The Journal of Neuroscience* 32: 7550-7562.

Yam, J. C. et al. (2019) Low-Concentration Atropine for Myopia Progression (LAMP) Study: A Randomized, Double-Blinded, Placebo-Controlled Trial of 0.05%, 0.025%, and 0.01% Atropine Eye Drops in Myopia Control. *Ophthalmology* 126:113-124.

Yang J, Lee S H, Wray N R, Goddard M E, and Visscher P M. (2016) GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. *Proceedings of the National Academy of Sciences* 113: e4579.

Yang J, Weedon M N, Purcell S, Lettre G, Estrada K, Willer C J, Smith A V et al. (2011a) Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics* 19: 807-812.

Yang J, Zaitlen N A, Goddard M E, Visscher P M, and Price A L (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* 46: 100-106.

Yang J A, Benyamin B, McEvoy B P, Gordon S, Henders A K, Nyholt D R, Madden P A et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42: 565–569.

Yang J A, Lee S H, Goddard M E, and Visscher P M (2011b) GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics* 88: 76-82.

Yang Z, Xiao X, Li S, and Zhang Q. (2009) Clinical and linkage study on a consanguineous Chinese family with autosomal recessive high myopia. *Molecular Vision* 15: 312-318.

Yazar S, Hewitt A W, Black L J, McKnight C M, Mountain J A, Sherwin J C, Oddy W H et al. (2014) Myopia is associated with lower vitamin D status in young adults. *Investigative Ophthalmology and Visual Science* 55: 4552-4559.

You Q S, Wu L J, Duan J L, Luo Y X, Liu L J, Li X, Gao Q et al. (2014) Prevalence of myopia in school children in greater Beijing: the Beijing Childhood Eye Study. *Acta Ophthalmolgica* 92: 398-406.

Young A I. (2019) Solving the missing heritability problem. *PLoS Genetics* 15: e1008222.

Young F A, Leary G A, Baldwin W R, West D C, Box R A, Harris E, and Johnson C. (1969) The transmission of refractive errors within eskimo families. *Optometry and Vision Science* 46: 676-685.

Young T L, Atwood L D, Ronan S M, Dewan A T, Alvear A B, Peterson J, Holleschau A et al. (2001) Further refinement of the MYP2 locus for autosomal dominant high myopia by linkage disequilibrium analysis. *Ophthalmic Genetics* 22: 69-75.

Young T L, Deeb S S, Ronan S M, Dewan A T, Alvear A B, Scavello G S, Paluru P C et al. (2004) X-linked high myopia associated with cone dysfunction. *Archives of Ophthalmology* 122: 897-908.

Young T L, Ronan S M, Alvear A B, Wildenberg S C, Oetting W S, Atwood L D, Wilkin D J et al. (1998) A second locus for familial high myopia maps to chromosome 12q. *American Journal of Human Genetics* 63: 1419-1424.

Yuan Y, Zhang Z, Zhu J, He X, Du E, Jiang K, Zheng W et al. (2015) Responses of the ocular anterior segment and refraction to 0.5% tropicamide in chinese school-aged children of myopia, emmetropia, and hyperopia. *Journal of Ophthalmology* 2015: e612728.

Zadnik K. (2001) Association between night lights and myopia: True blue or a red herring? *Archives of Ophthalmology* 119: 146.

Zadnik K, Sinnott L T, Cotter S A, Jones-Jordan L A, Kleinstein R N, Manny R E, Twelker J D et al. (2015) Prediction of Juvenile-Onset Myopia. *JAMA Ophthalmology* 133: 683-689.

Zhang G, Bacelis J, Lengyel C, Teramo K, Hallman M, Helgeland O, Johansson S et al. (2015a) Assessing the causal relationship of maternal height on birth size and gestational age at birth: a mendelian randomization analysis. *PLoS Med* 12: e1001865.

Zhang Q, Guo X, Xiao X, Jia X, Li S, and Hejtmancik J F. (2005) A new locus for autosomal dominant high myopia maps to 4q22-q27 between D4S1578 and D4S1612. *Molecular Vision* 11: 554-560.

Zhang Q, Guo X, Xiao X, Jia X, Li S, and Hejtmancik J F. (2006) Novel locus for X linked recessive high myopia maps to Xq23-q25 but outside MYP1. *Journal of Medical Genetics* 43: e20.

Zhang Q, Li S, Xiao X, Jia X, and Guo X. (2007) Confirmation of a genetic locus for X-linked recessive high myopia outside MYP1. *Journal of Human Genetics* 52: 469-472.

Zhang X, Qu X, and Zhou X. (2015b) Association between parental myopia and the risk of myopia in a child. *Experimental and Therapeutic Medicine* 9: 2420-2428.

Zhang Y, Kurup P, Xu J, Carty N, Fernandez S M, Nygaard H B, Pittenger C et al. (2010) Genetic reduction of striatal-enriched tyrosine phosphatase (STEP) reverses cognitive and cellular deficits in an Alzheimer's disease mouse model. *Proceedings of the National Academy of Sciences of the United States of America* 107: 19014-19019.

Zhao F, Wu J, Xue A, Su Y, Wang X, Lu X, Zhou Z et al. (2013) Exome sequencing reveals CCDC111 mutation associated with high myopia. *Human Genetics* 132: 913-921.

Zhou W J, Zhang Y Y, Li H, Wu Y F, Xu J, Lv S, Li G et al. (2016) Five-year progression of refractive errors and incidence of myopia in school-aged children in western China. *Journal of Epidemiology* 26: 386-395.

Zhu Z, Bakshi A, Vinkhuyzen Anna A E, Hemani G, Lee Sang H, Nolte Ilja M, van Vliet-Ostaptchouk Jana V et al. (2015) Dominance genetic variation contributes little to the missing heritability for human complex traits. *The American Journal of Human Genetics* 96: 377-385.

Ziller M J, Gu H, Muller F, Donaghey J, Tsai L T, Kohlbacher O, De Jager P L et al. (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500: 477-481.

10 Appendices

The following appendices display the code and scripts used to conduct the data analyses in this thesis (where used). The appendices have been organised into separate sections for each experiment chapter. The code used has been written in black, notes describing what is being performed are presented in green, and variable names that allow for quick replacement in repetitive scripts are highlighted in red text (only done for noted areas).

10.1 Appendix A Analyses for Chapter 4, Experiment 1

```
=============
Creation of allele and weighted polygenic risk scores - Unix
=============

#!/bin/bash
#
#PBS -q serial
#PBS -l select=1:ncpus=1
#PBS -l walltime=2:00:00
#PBS -N plink_score2
#PBS -o plink_score2_out
#PBS -e plink_score2_err
#PBS -P PR300

# ------------------------------
# ALSPAC GWAS SCORE
# ------------------------------

module load plink/1.9c3

##weighted scores
plink \
  --bfile
/scratch/share_PR300/Neema/ALSPAC/yp_cream2017_hits_allchr \
#location of files specifying information on ALSPAC children
  --missing-code -9,0,NA,na \
  --maf 0.05 \
#to ensure no unreliable rare variants are used that may be
present
  --score 1 2 4 /scratch/share_PR300/Neema/ALSPAC/scorefile.txt
\
#location of variant effect weights to apply for ALSPAC
  --out /scratch/share_PR300/Neema/ALSPAC/weightedscores.out
#output file for analysis

##allele scores - comments as above
plink \
  --bfile
/scratch/share_PR300/Neema/ALSPAC/yp_cream2017_hits_allchr \
  --missing-code -9,0,NA,na \
  --maf 0.05 \
```

```
  --score 1 2
/scratch/share_PR300/Neema/ALSPAC/allelescorefile.txt \
  --out /scratch/share_PR300/Neema/ALSPAC/nonweightedscores.out

module unload plink/1.9c3

=======================
script for calculating prediction linear models - R
=======================

## read in calculated weighted scores
library(readr)
SNPSUM <- read_delim("~/weightedscores.txt",+     " ",
escape_double = FALSE, trim_ws = TRUE)

##read in phenotype measures and NMP file
nmpdata <- ALSPAC_parentalmyopia

names(SNPSUM)[1] <- "rupal_id_new"
names(SNPSUM)[2] <- "qlet"
names(nmpdata)[1] <- "rupal_id_new"
names(nmpdata)[2] <- "qlet"
#adjust names of columns for merging

SNPsumfull <- merge.data.frame(SNPSUM, nmpdata,
by=c("rupal_id_new","qlet"))
#merge files
SNPsumfull <-  SNPsumfull[!(is.na(SNPsumfull$NumMyopicParents) |
SNPsumfull$NumMyopicParents==""), ]
#remove missing information
shapiro.test(SNPsumfull$SCORE)
#data indicates normal
SNPsumfull$newSCORE <- ((SNPsumfull$SCORE -
mean(SNPsumfull$SCORE))) /sd(SNPsumfull$SCORE)
#standardise z-score for mean 0 sd 1

##read in allele scores
library(readr)
SNPALL <- read_delim("~/nonweightedscores.txt",+     " ",
escape_double = FALSE, trim_ws = TRUE)
nmpdata <- neema_parentalmyopia_2016_12_06
names(SNPALL)[1] <- "rupal_id_new"
names(SNPALL)[2] <- "qlet"
names(nmpdata)[1] <- "rupal_id_new"
names(nmpdata)[2] <- "qlet"
#all as above but for allele score

SNPALLfull <- merge.data.frame(SNPALL, nmpdata,
by=c("rupal_id_new","qlet"))
SNPALLfull <-  SNPALLfull[!(is.na(SNPALLfull$NumMyopicParents) |
SNPALLfull$NumMyopicParents==""), ]
#merge and remove all with missing NMP data
shapiro.test(SNPALLfull$SCORE)
#data indicates not normal

yr7data <- SNPsumfull[which(!is.na(SNPsumfull$Yr7_avMSE)),]
yr15data <- SNPsumfull[which(!is.na(SNPsumfull$Yr15_avMSE)),]
count(yr7data$NumMyopicParents)
```

```
count(yr15data$NumMyopicParents)
yr7data2 <- SNPALLfull[which(!is.na(SNPALLfull$Yr7_avMSE)),]
yr15data2 <- SNPALLfull[which(!is.na(SNPALLfull$Yr15_avMSE)),]
count(yr7data2$NumMyopicParents)
count(yr15data2$NumMyopicParents)
# identifying numbers in each prediction group for allele and
weighted

## identify which of the models is best for predicting
refractive error
Allelescorecalculation7 <- lm(yr7data2$Yr7_avMSE ~
yr7data2$SCORE)
Allelescorecalculation15 <- lm(yr15data2$Yr15_avMSE ~
yr15data2$SCORE)
Summary(Allelescorecalculation7)
Summary(Allelescorecalculation15)
#summaries give values for adj.rsq

weightscorecalculation7 <- lm(yr7data$Yr7_avMSE ~ yr7data$SCORE)
weightscorecalculation15 <- lm(yr15data$Yr15_avMSE ~
yr15data$SCORE)
Summary(weightscorecalculation7)
Summary(weightscorecalculation15)
#summaries give values for adj.rsq - values higher than allele
score

anova(weightscorecalculation7, allelescorecalculation7)
anova(weightscorecalculation15, allelescorecalculation15)
#calculate significance of differences between allele and z
score PRS

#calculation of model fits 7 and 15
nmpyr7model <- lm(yr7data$Yr7_avMSE ~ yr7data$NumMyopicParents)
alleleyr7model <- lm(yr7data$Yr7_avMSE ~ yr7data$SCORE)
combinedmodel7 <- lm(yr7data$Yr7_avMSE~yr7data$NumMyopicParents
+ yr7data$SCORE +
yr7data$NumMyopicParents:yr7data$SCORE)

nmpyr15model <- lm(yr15data$Yr15_avMSE ~
yr15data$NumMyopicParents)
alleleyr15model <- lm(yr15data$Yr15_avMSE ~ yr15data$SCORE)
combinedmodel15 <-
lm(yr15data$Yr15_avMSE~yr15data$NumMyopicParents +
yr15data$SCORE + yr15data$NumMyopicParents:yr15data$SCORE)

library(psychometric)
mysum <- summary(modelx) #delete and replace modelx as needed
CI.Rsq(rsq=mysum$r.squared, n=(mysum$df[1]+mysum$df[2]), k=mysum
#calculate confidence int. for the models

#comparison of age 7 to age 15 in children who attended both
Comparisondata <- merge.data.frame(yr7data, yr15data
by=c("rupal_id_new","qlet"))

nmpyr7modelcomp <- lm(Comparisondata$Yr7_avMSE ~
Comparisondata$NumMyopicParents)
alleleyr7modelcomp <- lm(Comparisondata$Yr7_avMSE ~
Comparisondata$SCORE)
```

```
combinedmodel7comp <-
lm(Comparisondata$Yr7_avMSE~Comparisondata$NumMyopicParents +
Comparisondata$SCORE)
#recalculate scores for combined subset of children

nmpyr15modelcomp <- lm(Comparisondata$Yr15_avMSE ~
Comparisondata$NumMyopicParents)
alleleyr15modelcomp <- lm(Comparisondata$Yr15_avMSE ~
Comparisondata$SCORE)
combinedmodel15comp <-
lm(Comparisondata$Yr7_avMSE~Comparisondata$NumMyopicParents +
Comparisondata$SCORE)
#as above, but for age 15

============================
Linear Mixed Model analysis - R
============================

rm(list=ls())
library(nlme)
library(reshape)
library(ggplot2)
#load up packages as needed

data1 <- neema_parentalmyopia_2016_12_06
data2 <- SCOREfile
#load data
data1$visits <- 5 - is.na(data1$Yr7_avMSE) -
is.na(data1$Yr10_avMSE) - is.na(data1$Yr11_avMSE) -
is.na(data1$Yr12_avMSE) - is.na(data1$Yr15_avMSE)
#set out how many visits people attended
data1.5 <- subset(data1, (!is.na(data1[,5])))
data1.75 <- data1[which(data1$NumMyopicParents>=0,
!is.na(TRUE)),]
#remove those with missing data or no attendance

#first lmm for GRS
names(data2)[1] <- "rupal_id_new"
names(data2)[2] <- "qlet"
data2$PHENO <- NULL
data2$CNT   <- NULL
data2$CNT2  <- NULL
#organise files as needed
data3 <- merge(data2,data1.5,by=c("rupal_id_new","qlet"))
data4 <- data3[which(data3$visits >= 3),]
n    <- dim(data4)[1]
#merge and count the number that have over 3 attendances

data5 <- melt(data4, id.vars = c("rupal_id_new"), measure.vars =
c("Yr7_avMSE","Yr10_avMSE","Yr11_avMSE","Yr12_avMSE","Yr15_avMSE
"))
data6 <- melt(data4, id.vars =
c("rupal_id_new","sex","SCORE","visits"), measure.vars =
c("age7","age10","age11","age12","age15"))
#melt data to allow LMM to take place
names(data5)[2] <- "avMSE"
names(data5)[3] <- "avMSE_at_visit"
names(data6)[5] <- "age"
```

```
names(data6)[6] <- "age_at_visit"
#rename the columns for easier understanding
data7 <- data5[order(data5$rupal_id_new),]
data8 <- data6[order(data6$rupal_id_new),]
data7$visit <- rep(1:5,n)
data8$visit <- rep(1:5,n)
#order and organise file for LMM
data9 <- merge(data7,data8,by=c("rupal_id_new","visit"))
#merge data
mod1 <- lme(avMSE_at_visit ~ sex + SCORE + poly(I(age_at_visit -
7.53),4) + SCORE:I(age_at_visit - 7.53),
        random=~I(age_at_visit - 7.53) | rupal_id_new,
        correlation = corCAR1(form = ~ visit | rupal_id_new),
        na.action = na.omit,
        method="ML",
        data=data9)
summary(mod1)
#create first model

a <- summary(mod1)
table2 <- a$tTable
table2
#find coefficients and place in table
meann <- mean(data9$SCORE)
ci <- sd(data9$SCORE)
#calculate mean and stan dev of PRS
minscore  <- min(data9$SCORE)
fifth  <- meann-2*ci
lowrisk <- data9[which(data9$SCORE < mean-ci),]
highrisk <- data9[which(data9$SCORE > mean+ci),]
avrisk <- data9[which(data9$SCORE > mean-ci & data9$SCORE <
mean+ci),]
maxscore  <- max(data9$SCORE)
medscore  <- median(data9$SCORE)
#identify other values for PRS


pdata     <-expand.grid(age_at_visit=seq(7,15,by=1),
SCORE=c(highrisk,avrisk,lowrisk), sex=c(1.5),
avMSE_at_visit="1")
pdata[,4] <-predict(mod1, pdata, level=0)
#expand and apply lmm model data

pdata$SCORE <- as.factor(pdata$SCORE)
names(pdata)[2] <- "GeneticRiskScore"
levels(pdata$GeneticRiskScore) <- c("High Risk", "Average Risk",
"Low Risk.")
#restructure PRS LMM for preparation of figure

fig1 <- ggplot(pdata, aes(age_at_visit, avMSE_at_visit))  +
                labs (x="Age (Years)", y=NULL, linetype="Genetic
Risk Score",colour="Genetic Risk Score")  +
                scale_linetype_manual(values=c(1,1,1)) +
                geom_line  (size=1.5,
aes(linetype=pdata$GeneticRiskScore, colour=GeneticRiskScore),
show.legend = F) +
                geom_point (size=4,
aes(colour=GeneticRiskScore)) +
```

```
                #geom_errorbar(aes(ymin=meann-ci, ymax=mean+ci),
width=.1) +
                scale_shape_manual(values = c(16,16,16))  +
                scale_x_continuous(limits=c(7, 15.5),
breaks=seq(7, 15, 1)) +
                scale_y_continuous(limits=c(-1.5, 0.60),
breaks=seq(-1.5,0.5,0.25)) +
                scale_colour_manual(values=c("lightskyblue",
"green", "red")) +
                scale_fill_manual(values=  c("#FFFFFF",
"#00D800", "#000099", "#009E73", "#F0E442", "#0072B2",
"#D55E00", "#D55E00","#CC79A7")) +
                theme(legend.position=c(0.2,0.15),
axis.text=element_text(size=12),axis.title=element_text(size=14,
face="bold"), panel.grid.major = element_blank(),
panel.grid.minor = element_blank(),panel.background =
element_blank(), axis.line = element_line(colour = "black")) +
guides(colour = guide_legend(reverse=T) + theme(legend.key.width
= unit(2, "cm")))
fig1
#create LMM figure


#second lmm - noting not repeated in areas seen before

names(data2)[1] <- "rupal_id_new"
names(data2)[2] <- "qlet"
data2$PHENO <- NULL
data2$CNT   <- NULL
data2$CNT2  <- NULL

data3 <- merge(data2,data1.5,by=c("rupal_id_new","qlet"))
data4 <- data3[which(data3$visits >= 3),]
n    <- dim(data4)[1]

data5 <- melt(data4, id.vars = c("rupal_id_new"), measure.vars =
c("Yr7_avMSE","Yr10_avMSE","Yr11_avMSE","Yr12_avMSE","Yr15_avMSE
"))
data6 <- melt(data4, id.vars =
c("rupal_id_new","sex","NumMyopicParents","visits"),
measure.vars = c("age7","age10","age11","age12","age15"))
names(data5)[2] <- "avMSE"
names(data5)[3] <- "avMSE_at_visit"
names(data6)[5] <- "age"
names(data6)[6] <- "age_at_visit"
data7 <- data5[order(data5$rupal_id_new),]
data8 <- data6[order(data6$rupal_id_new),]
data7$visit <- rep(1:5,n)
data8$visit <- rep(1:5,n)
data9 <- merge(data7,data8,by=c("rupal_id_new","visit"))
head(data9)

mod1 <- lme(avMSE_at_visit ~ sex + NumMyopicParents +
poly(I(age_at_visit - 7.53),4) + NumMyopicParents:I(age_at_visit
- 7.53),
            random=~I(age_at_visit - 7.53) | rupal_id_new,
            correlation = corCAR1(form = ~ visit |
rupal_id_new),
```

```
            na.action = na.omit,
            method="ML",
            data=data9)
summary(mod1)

a <- summary(mod1)
table2 <- a$tTable
table2


minscore  <- 0
maxscore  <- 2
medscore  <- 1
#done to reflect the categorical natures of NMP

pdata     <-expand.grid(age_at_visit=seq(7,15,by=1),
NumMyopicParents=c(minscore,medscore,maxscore), sex=c(1.5),
avMSE_at_visit="1")
pdata[,4] <-predict(mod1, pdata, level=0)

pdata$NumMyopicParents <- as.factor(pdata$NumMyopicParents)
levels(pdata$NumMyopicParents) <- c("Zero","One","Two")

fig <- ggplot(pdata, aes(age_at_visit, avMSE_at_visit))  +
  labs (x="Age (Years)",y="Refractive Error (D)",
linetype="Number of Myopic Parents",colour="Number of Myopic
Parents")  +
  scale_linetype_manual(values=c(1,2,3)) +
  geom_line  (size=1.5, aes(linetype=pdata$NumMyopicParents),
show.legend = T) +
  geom_point (size=4) +
  #geom_errorbar(aes(ymin=meann-ci, ymax=mean+ci), width=.1) +
  scale_shape_manual(values = c(16,16,16))  +
  scale_x_continuous(limits=c(7, 15), breaks=seq(7, 15, 1)) +
  scale_y_continuous(limits=c(-1.50, 0.60), breaks=seq(-
1.50,0.6,0.25)) +
  scale_colour_manual(values=c("navyblue", "darkred", "purple"))
+
  scale_fill_manual(values=  c("#FFFFFF", "#00D800", "#000099",
"#009E73", "#F0E442", "#0072B2", "#D55E00",
"#D55E00","#CC79A7")) +
  theme(legend.position=c(0.28,0.15),
axis.text=element_text(size=12),axis.title=element_text(size=14,
face="bold"), panel.grid.major = element_blank(),
panel.grid.minor = element_blank(), panel.background =
element_blank(), axis.line = element_line(colour = "black")) +
guides(colour = guide_legend(reverse=F)) +
theme(legend.key.width = unit(2, "cm"))
fig

#3rd combined lmm - noting not done in areas seen before

names(data2)[1] <- "rupal_id_new"
names(data2)[2] <- "qlet"
data2$PHENO <- NULL
data2$CNT   <- NULL
data2$CNT2  <- NULL
```

```
data3 <- merge(data2,data1.5,by=c("rupal_id_new","qlet"))
data4 <- data3[which(data3$visits >= 3),]
n    <- dim(data4)[1]
#data4$combined <- lm(data4$avMSE_at_visit ~
data4$NumMyopicParents+data9$SCORE)
data5 <- melt(data4, id.vars = c("rupal_id_new"), measure.vars =
c("Yr7_avMSE","Yr10_avMSE","Yr11_avMSE","Yr12_avMSE","Yr15_avMSE
"))
data6 <- melt(data4, id.vars =
c("rupal_id_new","sex","NumMyopicParents","SCORE","visits"),
measure.vars = c("age7","age10","age11","age12","age15"))
names(data5)[2] <- "avMSE"
names(data5)[3] <- "avMSE_at_visit"
names(data6)[6] <- "age"
names(data6)[7] <- "age_at_visit"
data7 <- data5[order(data5$rupal_id_new),]
data8 <- data6[order(data6$rupal_id_new),]
data7$visit <- rep(1:5,n)
data8$visit <- rep(1:5,n)
data9 <- merge(data7,data8,by=c("rupal_id_new","visit"))
head(data9)


mod1 <- lme(avMSE_at_visit ~ sex + NumMyopicParents + SCORE +
NumMyopicParents:SCORE + poly(I(age_at_visit - 7.53),4) +
NumMyopicParents:I(age_at_visit - 7.53) + SCORE:I(age_at_visit -
7.53) + NumMyopicParents:SCORE:I(age_at_visit - 7.53),
            random=~I(age_at_visit - 7.53) | rupal_id_new,
            correlation = corCAR1(form = ~ visit |
rupal_id_new),
            na.action = na.omit,
            method="ML",
            data=data9)


summary(mod1)
a <- summary(mod1)
table2 <- a$tTable
table2

meann <- mean(data9$SCORE)
ci <- sd(data9$SCORE)


lowrisk <- data9[which(data9$SCORE < mean-ci),]
highrisk <- data9[which(data9$SCORE > mean+ci),]
avrisk <- data9[which(data9$SCORE > mean-ci & data9$SCORE <
mean+ci),]

meanx <-1
maxx<- 2
min<- 0

pdata     <-expand.grid(age_at_visit=seq(7,15,by=1),
SCORE=c(lowrisk,averagerisk, highrisk),
NumMyopicParents=c(min,meanx,maxx), sex=c(1.5),
avMSE_at_visit="1")
pdata[,5] <-predict(mod1, pdata, level=0)
```

```
pdata$SCORE <- as.factor(pdata$SCORE)
names(pdata)[2] <- "GeneticRiskScore"

pdata$NumberofMyopicParents <-
as.factor((pdata$NumMyopicParents))
levels(pdata$NumberofMyopicParents) <- c("None","One","Two")
levels(pdata$GeneticRiskScore) <- c("High Risk","Average Risk",
"Low Risk")
#specify new levels for individuals to fit into with both
variables

fig2 <-ggplot(pdata, aes(age_at_visit, avMSE_at_visit)) +
  labs (x="Age (Years)", y=NULL, linetype = "Number of Myopic
Parents", color="GeneticRiskScore") +
  scale_linetype_manual(values=c(1,2,3)) +
  geom_line  (size=1.5, aes(colour=GeneticRiskScore,
linetype=NumberofMyopicParents), show.legend = T) +
  geom_point (size=4, aes(colour=GeneticRiskScore,
fill=GeneticRiskScore)) +
  scale_shape_manual(values = c(21,21,21))  +
  scale_x_continuous(limits=c(7,15.5), breaks=seq(7, 15, 1)) +
  scale_y_continuous(limits=c(-1.50, 0.6), breaks=seq(-
1.75,0.75,0.25)) +
  theme(axis.text.x = element_blank() ) +
  theme(legend.position=c(0.25,0.22),panel.grid.major =
element_blank(), panel.grid.minor =
element_blank(),panel.background = element_blank(), axis.line =
element_line(colour = "black"),
axis.text.x=element_text(size=12),
axis.text.y=element_text(size=12),axis.title=element_text(size=1
4,face="bold")) + guides(colour = guide_legend(reverse=F)) +
theme(legend.key.width = unit(2, "cm")) +
  #annotate("text", x = 15, y=0.25,label = "a",cex=11, hjust =
0.5) +

scale_colour_manual(values=c("lightskyblue","green","red","light
skyblue","green","red","lightskyblue","green","red")) +
  scale_fill_manual(values=c("#000000", "#FFFFFF", "#000099",
"#009E73", "#F0E442", "#0072B2", "#D55E00",
"#D55E00","#CC79A7"))
fig2

======================
survival analysis scripts - R
======================
rm(list=ls())

library(ggplot2)
library(dplyr)
library(survival)
library(survminer)
#load up required packages

myp_threshold <- (-1.00)
#set myopia threshold
data0 <- neema_parentalmyopia_2016_12_06
data1 <- cream2017_ukbb_rep_profile_2018_01_02
```

```
#load up the data
names(data1)[1] <- "rupal_id_new"
names(data1)[2] <- "qlet"
data2       <-
merge.data.frame(data0,data1,by=c("rupal_id_new","qlet"))
#merge genetic and phenotype/NMP data
data2$event7     <- ifelse(data2$Yr7_avMSE  <= myp_threshold, 1,
0)
data2$event10    <- ifelse(data2$Yr10_avMSE <= myp_threshold, 1,
0)
data2$event11    <- ifelse(data2$Yr11_avMSE <= myp_threshold, 1,
0)
data2$event12    <- ifelse(data2$Yr12_avMSE <= myp_threshold, 1,
0)
data2$event15    <- ifelse(data2$Yr15_avMSE <= myp_threshold, 1,
0)
#identify dates that someone became myopic
data2$evermyopic <-
ifelse((rowSums(data2[,c("event7","event10","event11","event12",
"event15")], na.rm=TRUE) > 0), 1, 0)
data2$evermyopic <- ifelse((is.na(data2$event7) &
is.na(data2$event10) & is.na(data2$event11) &
is.na(data2$event12) & is.na(data2$event15)), NA,
data2$evermyopic)
#to find out if someone ever became myopic and remove any
missing info participants

data2$firstmyopic  <- ifelse(data2$event15==1, data2$age15, NA)
data2$firstmyopic  <- ifelse(data2$event12==1, data2$age12,
data2$firstmyopic)
data2$firstmyopic  <- ifelse(data2$event11==1, data2$age11,
data2$firstmyopic)
data2$firstmyopic  <- ifelse(data2$event10==1, data2$age10,
data2$firstmyopic)
data2$firstmyopic  <- ifelse(data2$event7==1,  data2$age7,
data2$firstmyopic)
#to correctly allocate the onset of myopia in the sample

data2$lastnevermyopic <- ifelse((data2$event7==0  &
data2$evermyopic==0), data2$age7, NA)
data2$lastnevermyopic <- ifelse(is.na(data2$event10),
data2$lastnevermyopic, ifelse((data2$event10==0 &
data2$evermyopic==0), data2$age10, data2$lastnevermyopic))
data2$lastnevermyopic <- ifelse(is.na(data2$event11),
data2$lastnevermyopic, ifelse((data2$event11==0 &
data2$evermyopic==0), data2$age11, data2$lastnevermyopic))
data2$lastnevermyopic <- ifelse(is.na(data2$event12),
data2$lastnevermyopic, ifelse((data2$event12==0 &
data2$evermyopic==0), data2$age12, data2$lastnevermyopic))
data2$lastnevermyopic <- ifelse(is.na(data2$event15),
data2$lastnevermyopic, ifelse((data2$event15==0 &
data2$evermyopic==0), data2$age15, data2$lastnevermyopic))
data3 <- data2[,c("rupal_id_new","sex", "NumMyopicParents",
"SCORE", "visits","evermyopic","time")]
data4 <- data3[which(data3$visits > 0 &
!is.na(data3$NumMyopicParents)),]
#finalise and subset individuals who have refractive data, NMP
data and have correct accurate myopia data
```

```
data4$newscore <- ((data4$SCORE -
mean(data4$SCORE))*1)/sd(data4$SCORE)
#z score standardise PRS
data4$NumMyopicParents <- as.factor(data4$NumMyopicParents)
#state that NMP is a factor

#meanscore <- mean(data4$SCORE)
#highci <- meanscore + 2*sd(data4$SCORE)
#lowci <- meanscore - 2*sd(data4$SCORE)
#serve as reminders
data4$GeneticRisk <- NULL
#create new column for grouping
data4$GeneticRisk <- ifelse(data4$newscore >= -1 &
data4$newscore <= 1, 2, 0)
data4$GeneticRisk <- ifelse(data4$newscore <= -1,
1,data4$GeneticRisk)
data4$GeneticRisk <- ifelse(data4$newscore >= 1, 3,
data4$GeneticRisk)
#subset the individuals into high med and low genetic risk
categories
data4$GeneticRisk <- as.factor(data4$GeneticRisk)
#change to factor to allow for levels in survival
levels(data4$GeneticRisk) <- c('Low Risk', 'Average Risk', 'High
Risk')

data4$NumMyopicParents <- as.integer(data4$NumMyopicParents)
data4$combined <- data4$NumMyopicParents *10
#alter NMP to integer to allow manipulation of two grouped
categories
data4$combined <- data4$combined + data4$GeneticRisk
#add genetic risk to NMP risk to allow for different groups
ranged from these two variables
data4$combined <- as.factor(data4$combined)
#reaffirm combined score into a categorical variable
library(plyr)
data4$combined <- revalue(data4$combined,
c("1"="LowGR;NoMP","2"="AvGR;NoMP","3"="HighGR;NoMP","11"="LowGR
;OneMP","12"="AvGR;OneMP","13"="HighGR;OneMP","21"="LowGR;TwoMP"
,"22"="AvGR;TwoMP","23"="HighGR;TwoMP"))
#label the groups according to risk

mod1  <- coxph(Surv(time, evermyopic) ~ sex + NumMyopicParents +
newscore + NumMyopicParents:newscore, ties="breslow",
data=data4)
summary(mod1)
mod2  <- coxph(Surv(time, evermyopic) ~ sex + NumMyopicParents,
ties="breslow", data=data4)
summary(mod2)
mod3  <- coxph(Surv(time, evermyopic) ~ sex + newscore,
ties="breslow", data=data4)
summary(mod3)
#create models for different survival analyses

mod2  <- survfit(Surv(time, evermyopic) ~ NumMyopicParents,
data=data4)
plot1 <- ggsurvplot(mod2, data=data4, xlim = c(8.5,15),
break.x.by =1, palette = c("black", "black", "black"),linetype =
c(1,2,3), censor=FALSE, legend=c(0.3,0.15), legend.title="Number
```

```
of myopic parents", legend.labs=c("Zero","One","Two"), xlab="Age
(years)", ylab="Proportion remaining non-myopic")
plot1
#create NMP plot
mod3  <- survfit(Surv(time, evermyopic) ~ GeneticRisk,
data=data4)
plot2 <- ggsurvplot(mod3, data=data4, xlim = c(8.5,15),
break.x.by =1, censor=FALSE, legend=c(0.25,0.15),
legend.labs=c("High Risk","Average Risk","Low Risk"),
legend.title="Genetic Risk Score", xlab="Age (years)", ylab=NULL
+theme(legend.key.width = unit(2, "cm")) + guide_legend(reverse
= TRUE))
plot2
#create PRS plot
mod4  <- survfit(Surv(time, evermyopic) ~ combined, data=data4)
plot3 <- ggsurvplot(mod4, data=data4, xlim = c(8.5,15),
break.x.by =1, linetype = c(1,1,1,2,2,2,3,3,3), palette =
c("red", "green","steelblue2", "red",
"green","steelblue2","red","green","steelblue2"),  censor=FALSE,
legend=c(0.36,0.23), legend.title="Combined Analysis",
legend.labs=c("High Genetic Risk; No Myopic Parents","Average
Genetic Risk; No Myopic Parents","Low Genetic Risk; No Myopic
Parents","High Genetic Risk; One Myopic Parent","Average Genetic
Risk; One Myopic Parent","Low Genetic Risk; One Myopic
Parent","High Genetic Risk; Two Myopic Parents","Average Genetic
Risk; Two Myopic Parents","Low Genetic Risk; Two Myopic
Parents"), legend.key.width=unit(5, "cm"), xlab="Age (years)",
ylab =NULL)
plot3
#create combined plot
```

10.2 Appendix B Analyses for Chapter 5, Experiment 2

```
==================================
Filtering outliers and quality control purposes for GWAS -
Unix
==================================

dataset <- ukb_european

#PCA filtering done as per bycroft et al. therefore other
quality control filters applied to sample
dataset <- dataset[which(!is.na(dataset$Sex_matched)),]
dataset <- dataset[which(!is.na(dataset$Age)),]

#below is for participants after heteozygosity filtering
meaneuro <- mean(dataset$Heterozygosity)
sdeuro <- sd(dataset$Heterozygosity)
datasetheterozycorrected <- dataset
[which(dataset$Heterozygosity <(meaneuro+(4*sdeuro))
& dataset$Heterozygosity >(meaneuro-(4*sdeuro))),]
#remove these last two lines if you want to use people for
predicted phenotypes and remove ! to allow only NA
dataset <- dataset[which(dataset$outlierMSE == 0),]
dataset <- dataset[which(!is.na(dataset$avMSE)),]
```

```
write.table(datasetheterozycorrected, file =
"avMSEphenoeuropean", sep = " ", row.names = FALSE,  append =
FALSE, quote = FALSE)
#write the file that includes participants for analysis

=======================
Bash script example for BOLT GWAS in autorefraction MSE - Unix
n = chromosome number of interest, change as needed
=======================
#!/bin/bash
#
#PBS -q workq
#PBS -l select=1:ncpus=8:mem=81GB
#PBS -l walltime=40:00:00
#PBS -N bolt
#PBS -o bolt_fullout
#PBS -e bolt_fullerrorr1
#PBS -P PR300
#
cd /scratch/share_PR300/Neema/Fullukbiobankwork/Boltscripts
#
#chr="n"
# ------------------------------
# Full dataset for Chr'n' UKBB GWAS
# ------------------------------
./bolt \
#call up software
 --
bgenFile=/scratch/share_PR300/EGA/imputed/ukb_imp_chrn_v2.bgen \
 --
bfile=/scratch/share_PR300/EGA/imputed/bolt/ukb_bolt_high_conf_h
rc_r0-05 \
 --sampleFile=/scratch/share_PR300/EGA/imputed/ukb_imp_v2.sample
\
 --
LDscoresFile=/scratch/share_PR300/Neema/Fullukbiobankwork/Tables
/LDSCORE.1000G_EUR.tab.gz \
 --
geneticMapFile=/scratch/share_PR300/Neema/Fullukbiobankwork/Tabl
es/genetic_map_hg19.txt.gz \
 --
phenoFile=/scratch/share_PR300/Neema/Fullukbiobankwork/phenofile
s/avMSEphenoeuropean \
#specify the files and location of data required for the BOLT
GWAS to be performed
 --lmm \
#LMM analysis
 --phenoCol=avMSE \
#name of phenotype
 --
remove=/scratch/share_PR300/Neema/Fullukbiobankwork/Tables/newre
movefile.remove \
#individuals to exclude (withdrawn etc)
 --maxMissingPerSnp=0.02 \
 --maxMissingPerIndiv=0.05 \
 --bgenMinMAF=0.01 \
 --bgenMinINFO=0.5 \
#QC measures
```

```
 --qCovarCol=Age \
 --qCovarCol=PC{1:20} -Unix \
#covariates to control for
 --
covarFile=/scratch/share_PR300/UKB/qc/covarianteuropeanfile.txt
\
 --covarCol=Geno_array \
 --covarCol=Sex_matched \
#specify covariate location and further covariates
 --noBgenIDcheck \
#flag to avoid checking IDs throughout file
 --
statsFile=/scratch/share_PR300/Neema/BOLT/chrnresults_250118.out
\
 --
statsFileBgenSnps=/scratch/share_PR300/Neema/BOLT/chrnresults_bg
en_250118.out\
#state output file locations for GWAS analysis

==============================
Script to merge all chromosomes into one file post analysis
- Unix
==============================

myfolder="/scratch/share_PR300/Neema/BOLT/avMSEresults"
#state folder location for work
# if MAF>0.01 and HWE -log10P<6 and INFO>0.5 then retain
#
echo "snpID rsID chr pos genpos A1 A0 A1freq INFO beta se P_bolt
majorAllele MAF HWE_log10P" >
${myfolder}/cat_chr_all_17052017.out
#state headings for new file to be created and the filters used
awk '{$12=$2":"$3"_"$5"_"$6};{print $0}'
/scratch/share_PR300/Neema/BOLT/avMSEresults/chr1.out | sort -
k12b,12 > ${myfolder}/cat1_chr1.out
#take out relevantr information from chr1 file
awk '{print $1,$8,$15,$16}'
/scratch/share_PR300/Neema/summary_stats_impv1_chr1.txt | sort -
k1b,1 > ${myfolder}/cat2_chr1.out
#take out relevant information from other files needed to carry
over
join -1 12 -2 1 ${myfolder}/cat1_chr1.out
${myfolder}/cat2_chr1.out > ${myfolder}/cat3_chr1.out
#join two files together having sorted both into order to match
awk '{if($14>0.01 && $15<6 && $9>0.5) print $0}'
${myfolder}/cat3_chr1.out >> ${myfolder}/cat_chr_all_dated.out
#apply QC filters

rm ${myfolder}/cat1_chr1.out
rm ${myfolder}/cat2_chr1.out
rm ${myfolder}/cat3_chr1.out
#remove temporary files

#as the headings were done via chr1, can loop the rest of the
information to be printed underneath through same filtering
process
for chr in {2..22}; do
```

```
awk '{$12=$2":"$3"_"$5"_"$6};{print $0}'
/scratch/share_PR300/Neema/BOLT/avMSEresults/chr${chr}.out |
sort -k12b,12 > ${myfolder}/cat1_chr${chr}.out
awk '{print $1,$8,$15,$16}'
/scratch/share_PR300/Neema/summary_stats_impv1_chr${chr}.txt |
sort -k1b,1 > ${myfolder}/cat2_chr${chr}.out
join -1 12 -2 1 ${myfolder}/cat1_chr${chr}.out
${myfolder}/cat2_chr${chr}.out > ${myfolder}/cat3_chr${chr}.out
awk '{if($14>0.01 && $15<6 && $9>0.5) print $0}'
${myfolder}/cat3_chr${chr}.out >>
${myfolder}/cat_chr_all_dated.out

rm ${myfolder}/cat1_chr${chr}.out
rm ${myfolder}/cat2_chr${chr}.out
rm ${myfolder}/cat3_chr${chr}.out

done

==================================
Script to make a manhattan plot and QQ plot - R
==================================
Library (qqman)
manhattan(cat_chr_all, col = c("red", "blue"))
#create manhattan plot with all variants
manhattan(cat_chr_all, col = c("red", "blue"), ymax = 20)
#create manhattan plot but with only those more than p x10-20
for visibility
qq(cat_chr_all$P_bolt)
#qqplot creation
```

10.3 Appendix C Analyses for Chapter 6, Experiment 3

```
==================================
Calculation for AOSW inferred MSE - R
==================================
data1 <- ukb_caucasian
#load data
data1$Sex <- as.factor(data1$Sex)
levels(data1$Sex) <- c("Male","Female")
#state sex as factor and rename for ease
data2 <- data1[which(!is.na(data1$avMSE)),]
data5 <- data1[which(is.na(data1$avMSE)),]
#create two sets for those with and without refractive data
respectively
data2$sample <- rbinom(n = dim(data2)[1], size = 1, p = 0.5)
datatest <- data2[which(data2$sample == 1),]
datatraining <- data2[which(data2$sample ==0),]
#divide sample with MSE data into two equal size datasets for
training and test
#Run Loop for polynomial of Age
threshold               <- 0
poly_count1             <- 0
graph_data              <-
as.data.frame(matrix(nrow=2,ncol=30))
names(graph_data)       <- c("Polynomial_order",
"Rsquared_of_model")
#create table and set out parameters to be used in loop
```

```
while(threshold < 0.05){
  poly_count1              <- poly_count1 + 1
  first_model              <- lm(avMSE ~ poly(Age, poly_count1)+
Sex, data=datatraining)
  second_model             <- lm(avMSE ~ poly(Age, poly_count1 +
1) + Sex, data=datatraining)
  stat_summary             <- anova(first_model, second_model)
  first_model_summary      <- summary(first_model)
  graph_data[poly_count1,1] <- poly_count1
  graph_data[poly_count1,2] <- first_model_summary$adj.r.squared
  threshold                <- unlist(stat_summary)[[12]]
}
#objective loop that prints to table the rsq and if it's more
significant to use higher order poly

# Now for AgeOnsetSpex - mostly as above, but different
parameter
threshold                <- 0
poly_count2              <- 0
graph_data               <-
as.data.frame(matrix(nrow=2,ncol=30))
names(graph_data)        <- c("Polynomial_order",
"Rsquared_of_model")

while(threshold < 0.05){
  poly_count1              <- poly_count2 + 1
  first_model              <- lm(avMSE ~ poly(AgeOnsetSpex,
poly_count2)+ Sex, data=datatraining)
  second_model             <- lm(avMSE ~ poly(AgeOnsetSpex,
poly_count2 + 1) + Sex, data=datatraining)
  stat_summary             <- anova(first_model, second_model)
  first_model_summary      <- summary(first_model)
  graph_data[poly_count2,1] <- poly_count2
  graph_data[poly_count2,2] <- first_model_summary$adj.r.squared
  threshold                <- unlist(stat_summary)[[12]]
}
calc1 <- lm(avMSE ~ poly(Age,poly_count1)+ Sex, data=datatest)
calc2   <-   lm(avMSE   ~   poly(AgeOnsetSpex,poly_count2)+   Sex,
data=datatest)
#calculations to work out the rsq of these in independent sample
combinedcalc <- lm(avMSE ~ poly(Age,poly_count1)+
poly(AgeOnsetSpex,poly_count2)+ Sex, data=datatest)
#work out rsq in combined sample
datatest$predicted_avMSE <- predict(combinedcalc,
data=datatest)
#created predicted MSE in datatest
rsq <- lm(avMSE ~ predicted_avMSE, data=datatest)
summary(rsq)
#recheck to see if rsq is the same with the predicted phenotype
datax$predicted_MSE <- predict(combinedcalc, data=data5)
#apply to all individuals with no rx after independent sample
validation
write.table(data5, file = "aoswMSEphenoeuropean", sep = " ",
row.names = FALSE,  append = FALSE, quote = FALSE)
#This is to be used in BOLT GWAS as pheno file
```

```
==================================
Script to inverse normalise transform to normal distribution - R
==================================

data2$ranked<-rank(data2$predicted_MSE, na.last = NA,
ties.method = c("first"))
#to rank according to predicted MSE

# new dataset for creating rank against.
data3 <- as.data.frame(matrix(nrow=98870,ncol=2))
data3[1] <- rnorm(n=98870, mean=0, sd=1)
#creation of the dataset with same no of rows, normal
distribution of values
names(data3)[1] <- "predictednormalised"
names(data3)[2] <- "ranked"
data3$ranked<-rank(data3$predictednormalised, na.last = NA,
ties.method = c("first"))
#rename columns to inform, rank them to allow merge
data4 <- merge.data.frame(data2, data3, by = "ranked")
#merging the dataframes together
newrsq <- lm(avMSE ~ predictednormalised, data=data4)
#rsq.adj reasonable, continue to use transformed dataset too for
estimating phenotype in non-mse individuals

# new normalised dataset for creating rank against.
data5 <- as.data.frame(matrix(nrow=287448,ncol=2))
data5[1] <- rnorm(n=287448, mean=0, sd=1)
names(data5)[1] <- "predictednormalised"
names(data5)[2] <- "ranked"
data5$ranked<-rank(data4$predictednormalised, na.last = NA,
ties.method = c("first"))
#same as above but to different sample size as required
data7 <- merge.data.frame(data6, data5, by = "ranked")
#merging the dataframes together, applied to all individuals
with no rx after validation in those with rx
write.table(data7, file = "aoswnormMSEphenoeuropean", sep = " ",
row.names = FALSE,  append = FALSE, quote = FALSE)
#This is to be used in BOLT GWAS as normalised pheno file


======================
Script to run Bolt GWAS analysis for AOSW inferred MSE and AOSW
normalised MSE - R
======================
#example of script for AOSW-inferred MSE and AOSW normalised MSE
GWAS, chromosomes and phenotype can change when needed

#for AOSW-inferred MSE chr n

#!/bin/bash
#
#PBS -q workq
#PBS -l select=1:ncpus=8:mem=91GB
#PBS -l walltime=30:00:00
#PBS -N bolt
```

```
#PBS -o bolt_fulloutpred
#PBS -e bolt_fullerrorpred1
#PBS -P PR300
#
cd /scratch/share_PR300/Neema/Fullukbiobankwork/Boltscripts
#file location
#
#chr="n"
# ------------------------------
# Full dataset for Chrn UKBB GWAS
# ------------------------------

./bolt \
#recall software
 --
bgenFile=/scratch/share_PR300/EGA/imputed/ukb_imp_chrn_v2.bgen \
 --
bfile=/scratch/share_PR300/EGA/imputed/bolt/ukb_bolt_high_conf_h
rc_r0-05 \
 --sampleFile=/scratch/share_PR300/EGA/imputed/ukb_imp_v2.sample
\
 --
LDscoresFile=/scratch/share_PR300/Neema/Fullukbiobankwork/Tables
/LDSCORE.1000G_EUR.tab.gz \
 --
geneticMapFile=/scratch/share_PR300/Neema/Fullukbiobankwork/Tabl
es/genetic_map_hg19.txt.gz \
 --
phenoFile=/scratch/share_PR300/Neema/Fullukbiobankwork/phenofile
s/aoswMSEphenoeuropean \
#as per avMSE, files and locations stated for GWAS for predicted
phenotypes
 --lmm \
 --phenoCol=predicted_avMSE \
 --
remove=/scratch/share_PR300/Neema/Fullukbiobankwork/Tables/newre
movefilepred.remove \
#remove exclusions/withdrawn
 --maxMissingPerSnp=0.02 \
 --maxMissingPerIndiv=0.05 \
 --bgenMinMAF=0.01 \
 --bgenMinINFO=0.5 \
#QC data
 --qCovarCol=Age \
 --qCovarCol=PC{1:20} \
 --covarFile=/scratch/share_PR300/UKB/qc/ukb_geno_mse2017-08-
29.txt \
 --covarCol=Geno_array \
 --covarCol=Sex_matched \
#covariate files and columns as per avMSE
 --noBgenIDcheck \
 --
statsFile=/scratch/share_PR300/Neema/BOLT/chrnresultspred_160218
.out\
 --
statsFileBgenSnps=/scratch/share_PR300/Neema/BOLT/chrnresultspre
d_bgen_160218.out\
#output files
```

```
#For prednorm in any chr 'n', relabel the phenotype file

#manipulate script from autorefraction MSE to merge all
chromosomes for AOSW inferred and AOSW normalised MSE into 1 file.
Same script to create manhattan plots

============================
Prepare and calculate genetic correlations - Unix
============================

#!/bin/bash
#
#PBS -q serial
#PBS -l walltime=00:06:00
#PBS -l select=1:ncpus=1:mem=10GB
#PBS -N geneticcor
#PBS -o geneticcorOUT
#PBS -e geneticcorERROR
#PBS -P PR300

cd
/scratch/share_PR300/Neema/BOLT/trueMSE/genetic_correlation/ldsc

module load ldsc/latest
module load python/2.7.4-ldsc
#load modules for software to run on RAVEN

# Munge Data - format files so they work with LDSC
./munge_sumstats.py \
--sumstats
/scratch/share_PR300/Neema/BOLT/trueMSE/genetic_correlation/ldsc
/pre_munge_chr_normpred \
#change to pre_munge_chr_pred or pre_munge_true for munging and
preparing other phenotypes
--snp snp \
--a1 a1 \
--a2 a2 \
--N 287448 \ #change value when performing for true
--p p \
--frq [colname] \
--signed-sumstats beta,0 \
#state column names for each aspect of the munging process
--out
/scratch/share_PR300/Neema/BOLT/trueMSE/genetic_correlation/ldsc
/fullprednormalised.munged \ #change for each phenotype output
--merge-alleles w_hm3.snplist
#output file for munged data

#now munged, use LDSC for GC

./ldsc.py \
#use LDSC
--ref-ld-chr eur_w_ld_chr/ \
```

```
#reference file for LD
--out threephenotypes.out \
#output files
--rg
fulltrue.munged.sumstats.gz,fullprednormalised.munged.sumstats.g
z,fullpred.munged.sumstats.gz \
--w-ld-chr eur_w_ld_chr/
#run genetic correlation comparisons between these files

module unload ldsc/latest
module unload python/2.7.4-ldsc
#unload modules used on RAVEN

================================
Scripts to assess correlation between predicted phenotypes
and autorefraction measured phenotypes - R
================================
preddata <- pre_munge_pred [,c("SNP", "BETA", "P")]
#change to pred normalised when needed
truedata <- pre_munge_actual [,c("SNP", "BETA", "P")]
colnames(preddata)[2] <- "PREDBETA"
colnames(preddata)[3] <- "PREDP"
colnames(realdata)[2] <- "REALBETA"
colnames(realdata)[3] <- "REALP"
combined <- merge.data.frame(preddata, realdata, by = "SNP")
#rename columns to allow merging of both files without confusion
library(ggplot2)

#test correlations at different p values which get smaller
throughout. Also plot the corresponding correlation
plot <- qplot(x = REALBETA, y = PREDBETA, data = combined, geom
= "point") + geom_smooth(method = "lm", colour = "red")
cor.test(combined$PREDBETA, combined$REALBETA)
combined1 <- combined [which(combined$REALP <= 0.5),]
combined1 <- combined [which(combined$PREDP <= 0.5),]
cor.test(combined$PREDBETA, combined$REALBETA)
plot0 <- qplot(x = REALBETA, y = PREDBETA, data = combined2,
geom = "point") + geom_smooth(method = "lm", colour = "red")
plot0
combined2 <- combined1 [which(combined$REALP <= 0.05),]
combined2 <- combined2 [which(combined2$PREDP <= 0.05),]
cor.test(combined2$PREDBETA, combined2$REALBETA)
plot1 <- qplot(x = REALBETA, y = PREDBETA, data = combined2,
geom = "point") + geom_smooth(method = "lm", colour = "red")
plot1
combined3 <- combined2 [which(combined2$REALP <= 0.005),]
combined3 <- combined3 [which(combined3$PREDP <= 0.005),]
cor.test(combined3$PREDBETA, combined3$REALBETA)
plot2 <- qplot(x = REALBETA, y = PREDBETA, data = combined3,
geom = "point")+ geom_smooth(method = "lm", colour = "red")
plot2
combined4 <- combined3 [which(combined3$REALP <= 0.0005),]
combined4 <- combined4 [which(combined4$PREDP <= 0.0005),]
cor.test(combined4$PREDBETA, combined4$REALBETA)
plot3 <- qplot(x = REALBETA, y = PREDBETA, data = combined4,
geom = "point")+ geom_smooth(method = "lm", colour = "red")
plot3
combined5 <- combined4 [which(combined4$REALP <= 0.00005),]
```

```
combined5 <- combined5 [which(combined5$PREDP <= 0.00005),]
cor.test(combined5$PREDBETA, combined5$REALBETA)
plot4 <- qplot(x = REALBETA, y = PREDBETA, data = combined4,
geom = "point") + geom_smooth(method = "lm", colour = "red")
plot4
#save all plotted correlations found
```

10.4 Appendix D Analyses for Chapter 7, Experiment 4

```
============================

Genetic correlations, the inclusion of education - Unix

============================

#munge then run through LDSC
module load ldsc/latest
module load python/2.7.4-ldsc

./ldsc.py \
--ref-ld-chr eur_w_ld_chr/ \
--out allphenotypes.out \
--rg
fulltrue.munged.sumstats.gz,fullprednormalised.munged.sumstats.g
z,fullpred.munged.sumstats.gz, fulledu.munged.sumstats.gz \
--w-ld-chr eur_w_ld_chr/

module unload ldsc/latest
module unload python/2.7.4-ldsc
#as before, load and run LDSC on munged sumstats, but include
education summary statistics to allow 4 way comparison


============================

Use of MTAG software for all correlated traits - Unix

============================

#mtag all correlated traits as required. Change traits as
necessary and add a third when needed, by replacing the MTAG trait
label

#!/bin/bash
#
#PBS -q workq
#PBS -l walltime=01:00:00
#PBS -l select=1:ncpus=1:mem=35GB
#PBS -N MTAGtrait1trait2
#PBS -o MTAGtrait1trait2
#PBS -e MTAGtrait1trait2err
#PBS -P PR300
#
cd /scratch/share_PR300/Neema/BOLT/trueMSE/mtag

module load python/2.7.11-genomics
module load ldsc/latest
#load modules needed to run MTAG software
```

```
./mtag.py \
#call software to use
 --sumstats
/scratch/share_PR300/Neema/BOLT/trueMSE/mtag/MTAGREADYTRAIT1.txt
,/scratch/share_PR300/Neema/BOLT/trueMSE/mtag/MTAGREADYTRAIT2.tx
t\
#recall the MTAG trait summary statistics to MTAG
 --out
/scratch/share_PR300/Neema/BOLT/trueMSE/mtag/trait1trait2files/t
rait1andtrait2\
#state output file

module unload python/2.7.11-genomics
module unload ldsc/latest
#unload modules from RAVEN


================================
Use of METAL for correlated traits. Comparison in meta-
analysis - Unix
================================

##script to run METAL software meta-analysis
./metalsoftware
#call software to use
SCHEME   STDERR
#state which form of metaanalysis - call for inverse variance form
MARKER   SNP
#variant names to meta anaylse
ALLELE   ALLELE1REF ALLELE2ALT
#allele labels
EFFECT   BETA
#effect size label
STDERR   SE
#Standard error label - needed for this meta-analysis method
PROCESS METALreadyTrait1
#first file name for meta-analysis

#for second file, labels stated as above, and file name given in
process. Note change METALreadyTrait2 to AOSW norm MSE when needed
MARKER   SNP
ALLELE   ALLELE1REF ALLELE2ALT
EFFECT   BETA
STDERR   SE
PROCESS METALreadyTrait2

ANALYZE
#command to run analysis

============================
Use of LDpred software to account for LD - Unix
============================

## LDpred software script. change trait name 'trait1' for
different trait or combined trait as required
```

```
#!/bin/bash
#
#PBS -l select=1:ncpus=1:mem=60GB
#PBS -l walltime=40:00:00
#PBS -N LDP_matrait
#PBS -o LDP_ma_outtrait
#PBS -e LDP_ma_errtrait
#PBS -P PR300

LD_validation_file="/scratch/share_PR300/Neema/Fullukbiobankwork
/LDvalidationfile.grm.id"
#file needed to create reference coordinate file
in_file="/scratch/share_PR300/Neema/BOLT/trait1/mtag/ldpred/ldpr
ed-master/ldpred/LDpredreadytrait1"
#summary statistics for trait of interest - trait 1
out_file="/scratch/share_PR300/Neema/BOLT/trait1/mtag/ldpred/ldp
red-master/ldpred/OUTFILEtrait1"
#state output file
cd /scratch/share_PR300/Neema/BOLT/trait1/mtag/ldpred/ldpred-
master/ldpred
#location to use workspace

#coordination first step to create and set up required LD
reference panel
./coord_genotypes.py \
 --gf ${LD_validation_file} \
 --vgf ${LD_validation_file} \
 --ssf ${in_file} \
 --N 95505 \
#change as per size of sample for each trait/combination
 --out ${out_file}-

#using coordination step files, start LDpred using the gibbs
sampler
./LDpred.py \
 --coord ${out_file} \
 --ld_radius 1000 \
 --num_iter 200 \
#state input options for gibbs sampler
 --local_ld_file_prefix ${out_file} \
 --N 95505 \
#as above, change when needed
 --out ${out_file}

##third step involes transferring out file from second step to
create PRS values. This has been done in plink

module load plink/1.9c3
#load plink software
plink \
  --bfile
/scratch/share_PR300/Neema/Fullukbiobankwork/validationsamplebfi
les \
#locate the files for information on the independent sample
wishing to create risk scores for
  --missing-code -9,0,NA,na \
  --score 2 3 6 ${out_file}\
```

```
#score file as found in LDpred second step
  --out /scratch/share_PR300/Neema/ALSPAC/ trait1outfilefinal
#outfile specification

module unload plink/1.9c3
#unload software from RAVEN

============================
Identifying significant differences between using lone and
combined traits for genetic prediction - R
============================
#Change name of trait 'trait1' as needed
data1 <- trait1file
data2 <- trait1andcombinationfile
#two risk score outputs to compare
linearmodel <- lm(data1$PHENO ~ data1$SCORE)
summary(linearmodel)
#to identify same file and same R2 as previous noted step

data1           <- data1[,c("IID","PHENO","SCORE")]
names(data1)    <- c("IID","true_phens","PRS_True")
data2           <- data2[,c("IID","SCORE")]
names(data2)    <- c("IID","PRS_TrueEdu")
#cut out unnecessary columns and rename
datax           <- merge(data1,data2,by="IID")
#merge the two models together
model1          <- lm(datax$true_phens ~ datax$trait1)
model2          <- lm(datax$true_phens ~ datax$traitcombination)
model3 <- lm(datax$true_phens ~ datax$trait1 +
datax$PRS_traitcombination)
anova(model2, model3)
#run linear regression models and anova to identify if models
significantly different. However not nested
datax$Diff      <- datax$PRS_traitcombination - datax$PRS_trait1
#identify the difference in the values
model5          <- lm(datax$true_phens ~
datax$PRS_traitcombinaion+ datax$Diff)
#create a model that accounts for difference between the
estimated risk scores in combined vs lone phenotypes
summary(model5)
anova(model1, model5, data=datax)
#run likelihood ratio test - identify if significant
library(psychometric)
mysum <- summary(model5)
#change model name as needed
CI.Rsq(rsq=mysum$r.squared, n=(mysum$df[1]+mysum$df[2]),
k=mysum$df[1], level = 0.95)
#calculate CI values

============================
Script example for AUC calculations - R
============================
#change the datafile and trait name as needed for each
calculation

datafortrait1 <- trait1ldpredoutput
#or mtag output if combined trait used
```

```
#ROC and corresponding AUC calculation for any myopia threshold
for genetic risk
datatrait1only$myopic <- as.numeric(datatrait1only$PHENO <= -
0.75)
model1 <- glm(myopic ~ datatrait1only$SCORE
,data=datatrait1only,family=binomial())
roc1 <- roc(response=model1$y, predictor=model1$fitted.values)
ci(model1)

#ROC and corresponding AUC calculation for moderate myopia
threshold for genetic risk
datatrait1only$myopic <- as.numeric(datatrait1only$PHENO <= -
3.00)
model1 <- glm(myopic ~ datatrait1only$SCORE
,data=datatrait1only,family=binomial())
roc2 <- roc(response=model1$y, predictor=model1$fitted.values)
ci(model1)

#ROC and corresponding AUC calculation for high myopia threshold
for genetic risk
datatrait1only$myopic <- as.numeric(datatrait1only$PHENO <= -
5.00)
model1 <- glm(myopic ~ datatrait1only$SCORE
,data=datatrait1only,family=binomial())
roc3 <- roc(response=model1$y, predictor=model1$fitted.values)
ci(model1)

#to identify any differences between AUROCs, use the following.
Change models to test as required
Roc.test(model1, model2, method =c("bootstrap"))

==============================
Script to compare different stratifications of genetic risk - R
==============================

data <- autorefractionaoswedufile
#should you want to change trait phenotype, can change file
input
linearmodel <- lm(data$PHENO ~ data$SCORE)
summary(linearmodel)
#to identify same file and getting same R2 as before!

data$rank <- rank(data$SCORE, na.last = NA, ties.method =
c("first"))
#create new variable for allowing a new dataset join that can be
ranked to identify the top 25,10, and 5% at risk
1516*0.25
1516*0.10
1516*0.05
#use these to identify and group individuals who are at highest
risk based on their relative location to these following ranks
datawithrank <- data

datawithrank$myopic <- ifelse(datawithrank$PHENO <=-0.75,1,0)
datawithrank$modmyopic <- ifelse(datawithrank$PHENO <=-3,1,0)
datawithrank$highmyopic <- ifelse(datawithrank$PHENO <=-5,1,0)
#label individuals with real refractive error who meet these
values
```

```
percent <- 10
#state the level of which ranked individuals are defined as high
risk
datawithrank$highrisk <- ifelse(datawithrank$rank >
quantile(datawithrank$rank, prob = 1 - percent/100),1,0)
#label those high risk
riskmodel1 <- glm(datawithrank$myopic ~ datawithrank$highrisk,
family = binomial())
#run glm for risk of any myopia
summary(riskmodel1)
exp(coef(riskmodel1))
(exp(cbind(OR=coef(riskmodel1), confint(riskmodel1))))
#identify odds ratio of being at risk when in this category

#same as above but for moderate risk
riskmodel2 <- glm(datawithrank$modmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel2)
exp(coef(riskmodel2))
(exp(cbind(OR=coef(riskmodel2), confint(riskmodel2))))

#as above but for high level risk
riskmodel3 <- glm(datawithrank$highmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel3)
exp(coef(riskmodel3))
(exp(cbind(OR=coef(riskmodel3), confint(riskmodel3))))

#as above, but for different percentile of risk
percent <- 5
#change threshold for labelling high risk
datawithrank$highrisk <- ifelse(datawithrank$rank >
quantile(datawithrank$rank, prob = 1 - percent/100),1,0)
riskmodel1 <- glm(datawithrank$myopic ~ datawithrank$highrisk,
family = binomial())
summary(riskmodel1)
exp(coef(riskmodel1))
(exp(cbind(OR=coef(riskmodel1), confint(riskmodel1))))

riskmodel2 <- glm(datawithrank$modmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel2)
exp(coef(riskmodel2))
(exp(cbind(OR=coef(riskmodel2), confint(riskmodel2))))

riskmodel3 <- glm(datawithrank$highmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel3)
exp(coef(riskmodel3))
(exp(cbind(OR=coef(riskmodel3), confint(riskmodel3))))

#as above for lower percentile of risk
percent <- 25

datawithrank$highrisk <- ifelse(datawithrank$rank >
quantile(datawithrank$rank, prob = 1 - percent/100),1,0)
```

```
riskmodel1 <- glm(datawithrank$myopic ~ datawithrank$highrisk,
family = binomial())
summary(riskmodel1)
exp(coef(riskmodel1))
(exp(cbind(OR=coef(riskmodel1), confint(riskmodel1))))

riskmodel2 <- glm(datawithrank$modmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel2)
exp(coef(riskmodel2))
(exp(cbind(OR=coef(riskmodel2), confint(riskmodel2))))

riskmodel3 <- glm(datawithrank$highmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel3)
exp(coef(riskmodel3))
(exp(cbind(OR=coef(riskmodel3), confint(riskmodel3))))
#translate odds ratios calculated into table along with their
corresponding Cis identified for all percentiles and thresholds
of risk
```

10.5 Appendix E Analyses for Chapter 8, Experiment 5

```
==============================
Script to remove participants not within +/- 10 SD mean PC
- R
==============================

totalparticipantlist <- ukb_phenotypefile_ethnicityxonly
#change ethnicityx as needed

#Loop for mean and SD of different PCs
PCA <- 1
column_number <-12
#starting column and PCA value for loop
results_table <-as.data.frame(matrix(nrow=20, ncol=3))
colnames(results_table) <- c("PCA","MEAN","SD")
#create table to transcribe PCs to
 for(loop in 1:10) {
   results_table[loop,1] <- PCA
   results_table[loop,2] <- mean(dataset[,column_number],
na.rm=TRUE)
   results_table[loop,3] <- sd(dataset[,column_number], na.rm =
TRUE)
   PCA<- PCA+1
   column_number<- column_number+1
}
#calculate mean and SDs for PC values for different PCs in a
loop as continuing
dataset$PCApass <- ifelse(((dataset$PCA{n} >
mean(dataset$PCA{n}) - 10*sd(dataset$PCA{n})) & ((dataset$PCA{n}
< mean(dataset$PCA{n}) + 10*sd(dataset$PCA{n}))1,0)
#change as needed for different PC values of n
newpasseddataset <- dataset[which(dataset$PCApass == 1),]
#filter out those that are within top 10 PC values for the self-
reported trait
```

```
write.table (newpasseddataset, file =”newethnicityxpcafiltered”
sep = " ", row.names = FALSE,  append = FALSE, quote = FALSE
#create file for ethnicity of interest to calculate risk scores


====================================
Script to calculate risk scores using PLINK - Unix
====================================

#as no change to mtag and phenotypes used just need to adapt
risk scores for these individuals that have been filtered.
change trait name for trait or traitconmbination as required, as
well as name and files for ethnicity of interest with ethnicityx

#!/bin/bash
#
#PBS -l select=1:ncpus=1:mem=60GB
#PBS -l walltime=40:00:00
#PBS -N LDP_matraitethnicityx
#PBS -o LDP_ma_outtraitethnicityx
#PBS -e LDP_ma_errtraitethnicityx
#PBS -P PR300

#load plink as before, and run software on new ethnicity
module load plink/1.9c3
plink \
  --bfile
/scratch/share_PR300/Neema/Fullukbiobankwork/ethnicityx/validati
onsampleethnicityxbfiles \
  --missing-code -9,0,NA,na \
  --score 2 3 6 newethnicityxfiltered \
#use ethnicity file created. Change for each ethnicity
  --out /scratch/share_PR300/Neema/Fullukbiobankwork/ethnicityx/
trait1outfilefinalethnicityx

module unload plink/1.9c3
#output file written and unload software from RAVEN

============================
Script to find if including a trait combination is better
than trait alone - R
============================

##calculate intra-ethnic differences in models used. based on
previous chapter likelihood ratio model. change trait and
ethnicityx as needed
data1 <- trait1ethnicityxfile
data2 <- trait1andcombinationethnicityxfile
linearmodel <- lm(data1$PHENO ~ data1$SCORE)
summary(linearmodel)
#to identify same file and getting same R2 as LDpred software

#script for analysis similar to previous chapter, change only to
input file from each ethnicity being tested
data1           <- data1[,c("IID","PHENO","SCORE")]
names(data1)    <- c("IID","true_phens","PRS_True")
data2           <- data2[,c("IID","SCORE")]
```

```
names(data2)    <- c("IID","PRS_TrueEdu")

datax           <- merge(data1,data2,by="IID")

model1          <- lm(datax$true_phens ~ datax$trait1)
model2          <- lm(datax$true_phens ~ datax$traitcombination)
model3 <- lm(datax$true_phens ~ data5$trait1 +
data5$PRS_traitcombination)
summary(model1)
summary(model2)
anova(model2, model3

datax$Diff      <- datax$PRS_traitcombination - datax$PRS_trait1
model5          <- lm(datax$true_phens ~
datax$PRS_traitcombinaion+ datax$Diff)
summary(model5)
anova(model1, model5, data=datax)
library(psychometric)
mysum <- summary(model5)#change model as needed
CI.Rsq(rsq=mysum$r.squared, n=(mysum$df[1]+mysum$df[2]),
k=mysum$df[1], level = 0.95)
```

==========

Script for ROC and corresponding AUC calculations - R

==========

```
dataforethnicityx <- autorefractionaosweduldpredoutput
#change as required for ethnicity of interest

#as previous analyses, compare each ethnicity AUC when looking
at best myopia prediction model
dataforethnicityx$myopic <- as.numeric(dataforethnicityx$PHENO
<= -0.75)
model1 <- glm(myopic ~ dataforethnicityx$SCORE
,data=dataforethnicityx,family=binomial())
roc1 <- roc(response=model1$y, predictor=model1$fitted.values)

dataforethnicityx$myopic <- as.numeric(dataforethnicityx$PHENO
<= -3.00)
model1 <- glm(myopic ~ dataforethnicityx$SCORE
,data=dataforethnicityx,family=binomial())
roc2 <- roc(response=model1$y, predictor=model1$fitted.values)

dataforethnicityx$myopic <- as.numeric(dataforethnicityx$PHENO
<= -5.00)
model1 <- glm(myopic ~ dataforethnicityx$SCORE
,data=dataforethnicityx,family=binomial())
roc3 <- roc(response=model1$y, predictor=model1$fitted.values)

#to identify any differences between AUROCs, use the following.
Change models to test as required
roc.test(model1, model2, method =c("bootstrap"))
```

==========

Script to compare different stratifications of genetic risk in different ethnicities - R

==========

```
#script for calculating different risks based on stratification,
as per previous chapter. Change of first couple of lines allows
script to run as per previous script

data <- autorefractionaoswedufilethnicityx
#should you want to change ethnicity, change the name of
ethnicityx
linearmodel <- lm(data$PHENO ~ data$SCORE)
summary(linearmodel)
#to identify same file and getting same R2 as software


data$rank <- rank(data$SCORE, na.last = NA, ties.method =
c("first"))
#create new variable for allowing a new dataset join that can be
ranked to identify the top 25,10, and 5% at risk
1516*0.25
1516*0.10
1516*0.05
#use these to identify and group individuals who are at highest
risk based on their relative location to these following ranks
datawithrank <- data

datawithrank$myopic <- ifelse(datawithrank$PHENO <=-0.75,1,0)
datawithrank$modmyopic <- ifelse(datawithrank$PHENO <=-3,1,0)
datawithrank$highmyopic <- ifelse(datawithrank$PHENO <=-5,1,0)

percent <- 10

datawithrank$highrisk <- ifelse(datawithrank$transformedgrs >
quantile(datawithrank$transformedgrs, prob = 1 -
percent/100),1,0)
riskmodel1 <- glm(datawithrank$myopic ~ datawithrank$highrisk,
family = binomial())
summary(riskmodel1)
exp(coef(riskmodel1))
(exp(cbind(OR=coef(riskmodel1), confint(riskmodel1))))

riskmodel2 <- glm(datawithrank$modmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel2)
exp(coef(riskmodel2))
(exp(cbind(OR=coef(riskmodel2), confint(riskmodel2))))

riskmodel3 <- glm(datawithrank$highmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel3)
exp(coef(riskmodel3))
(exp(cbind(OR=coef(riskmodel3), confint(riskmodel3))))

percent <- 5

datawithrank$highrisk <- ifelse(datawithrank$transformedgrs >
quantile(datawithrank$transformedgrs, prob = 1 -
percent/100),1,0)
riskmodel1 <- glm(datawithrank$myopic ~ datawithrank$highrisk,
family = binomial())
summary(riskmodel1)
```

```
exp(coef(riskmodel1))
(exp(cbind(OR=coef(riskmodel1), confint(riskmodel1))))

riskmodel2 <- glm(datawithrank$modmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel2)
exp(coef(riskmodel2))
(exp(cbind(OR=coef(riskmodel2), confint(riskmodel2))))

riskmodel3 <- glm(datawithrank$highmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel3)
exp(coef(riskmodel3))
(exp(cbind(OR=coef(riskmodel3), confint(riskmodel3))))


percent <- 25

datawithrank$highrisk <- ifelse(datawithrank$transformedgrs >
quantile(datawithrank$transformedgrs, prob = 1 -
percent/100),1,0)
riskmodel1 <- glm(datawithrank$myopic ~ datawithrank$highrisk,
family = binomial())
summary(riskmodel1)
exp(coef(riskmodel1))
(exp(cbind(OR=coef(riskmodel1), confint(riskmodel1))))

riskmodel2 <- glm(datawithrank$modmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel2)
exp(coef(riskmodel2))
(exp(cbind(OR=coef(riskmodel2), confint(riskmodel2))))

riskmodel3 <- glm(datawithrank$highmyopic ~
datawithrank$highrisk, family = binomial())
summary(riskmodel3)
exp(coef(riskmodel3))
(exp(cbind(OR=coef(riskmodel3), confint(riskmodel3))))
#as previously, results need to be transcribed to table with
corresponding CIs, and for different ethnicityx required
```