UNIVERSITY OF LEICESTER

Male and female IPF polygenic risk scores and predictive accuracy

Anne Florentine Goemans

MSc in Medical Statistics

University of Leicester

Department of Health Sciences

Submitted: 3rd January, 2022

SUMMARY

Idiopathic Pulmonary Fibrosis is a rare and devastating disease of the lungs. Within IPF there is a male predominance in cases. Recently sex differences in the genetic architecture of other traits have been identified. Establishing if such a sex difference exists in IPF could increase our understanding of the disease. Additionally, this project aims to test whether taking sex into account can help to improve prediction accuracy of a polygenic risk score (PRS).

A GRS of just the 15 known IPF risk variants had moderately high predictive accuracy with an overall AUC of 0.787 (0.768, 0.806), the sex difference was not significant. The predictive accuracy did not improve further for more complex analyses with more variants or sex-specific effect estimates.

A meta-analysis of female specific effect sizes from a sex-interaction GWAS did not identify any new variants with female specific genome-wide significance, but 16 variants of interest were identified including a variant in ADAM33, a gene which has been implicated in IPF in some previous studies.

More work needs to be done with larger datasets and sex specific GWAS before we can conclude whether there are sex differences within the genetics of IPF.

ACKNOWLEDGEMENTS

I'd like to thank my supervisors Professor Louise Wain and Dr Olivia Leavy for their guidance and discussions throughout this project.

Thanks to PPD for providing the financial support and flexibility to allow me to undertake this MSc.

Finally, thank you Peter, for keeping me sane. I think it's fair to say that I could not have done it without your support.

TABLE OF CONTENTS

1 IDIOPATHIC PULMONARY FIBROSIS BACKGROUND 1

1.1 Idiopathic Pulmonary Fibrosis 1

1.1.1 Clinical Signs & Natural History 1

1.1.2 Diagnosis 2

1.1.3 Treatment 2

1.1.4 Incidence and prevalence 3

1.1.5 Pathophysiology 4

1.1.6 Risk factors 4

1.2 Genetics of IPF 5

1.2.1 Basic Genetic Concepts 5

1.2.2 Overview of IPF Genetics 8

1.2.3 MUC5B 11

1.3 Sex differences in IPF 12

1.3.1 Differences in IPF traits 13

1.3.2 Factors associated with sex difference in IPF risk 13

1.4 Prediction models in IPF 16

1.5 Objectives and Outline 18

1.6 Summary 19

2 POLYGENIC RISK SCORES BACKGROUND 20

2.1 Polygenic Risk Scores 20

2.2 PRS methods 21

2.2.1 PRSice 22

2.3 Assessing performance and Predictive accuracy 23

2.3.1 Distribution of Score 23

2.3.2 Association 24

2.3.3 Discriminative ability and AUC....25
2.3.4 Variance explained....27
2.3.5 Risk Classification....29
2.3.6 Validation....29
2.4 Applications of PRS scores....29
2.5 Applications in IPF and this project....30
2.6 Summary....31
3 GENETIC RISK SCORE WITH KNOWN IPF RISK VARIANTS....32
3.1 Methods....32
3.1.1 Data....32
3.1.2 GRS calculation using R....34
3.1.3 GRS Association with phenotype....35
3.1.4 GRS Discriminative ability....36
3.1.5 GRS Clinical utility....36
3.1.6 Sensitivity – Effect of individual variants on GRS....37
3.1.7 Sensitivity – Using PRSice to create GRS....37
3.2 Results....38
3.2.1 GRS calculation using R....38
3.2.2 GRS Association with phenotype....40
3.2.3 GRS Discriminative ability....42
3.2.4 Clinical utility of GRS score....43
3.2.5 Sensitivity – Effect of individual variants on GRS....44
3.2.6 Sensitivity - using PRSice to generate GRS....45
3.3 Discussion....47
3.3.1 GRS using R....47
3.3.2 GRS using PRSice....48

3.4 Summary 49
4 PRS WITH STANDARD GWAS 50
4.1 Methods 50
4.1.1 Data 50
4.1.2 Data preparation 51
4.1.3 Thresholding of PRS scores 51
4.1.4 Assessment of performance 52
4.1.5 Optimising PRS score by sex 53
4.1.6 Sensitivity - Ambiguous SNP inclusion 54
4.1.7 Sensitivity – Minor Allele Frequency 54
4.1.8 Sensitivity – Model checking for outliers and influence 55
4.2 Results 55
4.2.1 Thresholding of PRS scores 55
4.2.2 Assessment of overall performance – PRSice model fit 58
4.2.3 Optimisation of sex differences 58
4.2.4 PRS optimised by sex 64
4.2.5 Sensitivity – Ambiguous SNPs 67
4.2.6 Sensitivity - Minor Allele Frequency 69
4.2.7 Outliers and Influence 70
4.3 Discussion 71
4.3.1 Sex differences 71
4.3.2 Sex-specific risk prediction 72
4.4 Summary 73
5 PRS WITH SEX INTERACTION GWAS 74
5.1 Methods 75
5.1.1 Data 75

5.1.2 GWA Meta-analysis....76
5.1.3 PRS scoring Sex Interaction....78
5.1.4 Sex-interaction effect direction to inform SNP selection....78
5.1.5 Sex-specific terms from interaction meta-analysis....79
5.2 Sex interaction term Results....80
5.2.1 Meta-analysis sex interaction term....80
5.2.2 PRS Scoring and Performance sex-interaction term and effects....82
5.2.3 Using Effect direction to inform SNP selection....84
5.2.4 Partitioned SNPs and the most predictive score by sex....86
5.3 Female specific SNP association results....87
5.3.1 Meta-analysis female specific SNP term....87
5.3.2 PRS Scoring and Performance female specific SNP term....90
5.3.3 GRS + female specific variants of interest....92
5.4 Discussion....93
5.4.1 Sex-interaction term assessment....93
5.4.2 Sex-interaction term direction to inform SNP selection....93
5.4.3 Female specific effect estimates....94
5.5 Summary....95
6 DISCUSSION....97
6.1 Overall Findings....97
6.2 Limitations and Further Research....98
6.2.1 GWAS effect estimates....98
6.2.2 Sex-Specific GWAS effect estimates....100
6.2.3 Bias in effect size estimates – Winner's curse....100
6.2.4 PRS method used....101
6.2.5 PRS Clinical utility....103

6.3 Overall Conclusions 104

REFERENCES 105

APPENDICES 112

APPENDIX A 112

APPENDIX B 115

APPENDIX C 117

APPENDIX D 118

APPENDIX E 119

LIST OF TABLES

Table 1.1: Known variants associated with risk of IPF. 11

Table 1.2: Clinical prediction models in IPF.....17

Table 3.1: UUS data case and control by sex.....38

Table 3.2: Overall summary statistics of the Genetic Risk Score 39

Table 3.3: Logistic regression of GRS against disease status model summary results.....41

Table 3.4: Predictive accuracy of GRS using Receiver Operator Characteristics 42

Table 3.5: PPV at specific GRS score thresholds and population prevalence estimates; [1] Percentile of score. 43

Table 3.6: Ambiguous SNPs removed from GRS by PRSice 45

Table 3.7: PRSice v1.25 compared to manual code results. [#]PRSice modified to allow for inclusion of ambiguous SNPs [1]p-value of score from model in equation (3.2) [2]p-value from de Long's test of 2 ROC curves for difference in discriminative ability of the score when applied to males and females separately 46

Table 4.1: Summary of Results from initial default PRS thresholds. [1]p-value from de Long's test of 2 ROC curves [2]Score p-value as generated by PRSice, from logistic regression including score + covariates [3]p-value from likelihood ratio test of model including sex interaction against base model 57

Table 4.2: Thresholds with largest difference in performance between the sexes [1]p-value from de Long's test of 2 ROC curves [2]Score p-value as generated by PRSice, from logistic regression including score + covariates [3]p-value from likelihood ratio test of model including sex interaction against base model 62

Table 4.3: Relative Performance of best Male and Female thresholds 65

Table 4.4: Proxies for Selected Ambiguous SNPs. LD based on all European reference populations available in LDLink 68

Table 4.5: Sensitivity summary - thresholds with largest difference in performance between the sexes [1]p-value from de Long's test of 2 ROC uncorrelated curves [2]Score p-value as generated by PRSice, from logistic regression including score + covariates [3]p-value from likelihood ratio test of model including sex interaction against base model.....71

Table 5.1: Summary of Results from initial default PRS thresholds. [1]p-value from de Long's test of 2 uncorrelated ROC curves [2]Score p-value as generated by PRSice, from logistic regression including score + covariates 83
Table 5.2: Performance of PRS scores, using sex interaction effect direction 87
Table 5.3: Possible hits for females from sex interaction GWAS, $p < 1x10^{-5}$ 89
Table 5.4: Relative Performance of best Male and Female thresholds 92
Table 5.5: Performance of GRS + female hits, compared to GRS using sex-agnostic and female specific effect sizes [1]p-value of score from model in equation (3.2) [3]SPDL1 variant missing from female effect estimates due to MAF < 1% 92

LIST OF FIGURES

Figure 1.1: Molecular structure of DNA, showing the 2 base pairs[25].....6
Figure 1.2: Chromosomal recombination during meiosis and inheritance from parents to offspring. Simplified example showing 1 chromosome.....7
Figure 1.3: Familial to Sporadic continuum of IPF, from Kropski, J.A., T.S. Blackwell, and J.E. Loyd(2015)[31].....8
Figure 1.4: IPF genetic variants by risk allele frequency and strength of genetic effect (OR). Genes/variants in white text primarily involved with familial IPF, genes in black important in both sporadic and familial. Note that the effect size and frequency is illustrative. Adapted from TA Manolio et al. Nature 461, 747-753 (2009)[36] 10
Figure 3.1: Boxplot of distribution of genetic risk score by disease status. Box represents the main body of the data, bound by the 25th and 75th percentile; outliers shown by open circles. 39
Figure 3.2: Genetic Risk Score distribution by disease status and sex for 15 variants with known IPF risk 40
Figure 3.3: ROC curves of Genetic Risk Score, overall and by sex.....42
Figure 3.4: PPV of the 15 SNP GRS score at selected prevalence. Red line = 1.25, black= 63 and blue=495 cases per 100,000 44
Figure 3.5: Effect of removing individual variants on performance of GRS.....45
Figure 4.1: Creation and testing of optimum scores by sex.....54
Figure 4.2: Model fit of low-resolution default thresholds. R^2 on the liability scale assuming a population prevalence of 63 cases per 100,000.....56
Figure 4.3: AUC by Sex across p-value thresholds, low-resolution scoring. Linear x-axis (right); Log x-axis (left). Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI. 58
Figure 4.4: PRS model fit across p-value thresholds. Black line shows model fit with score and covariates; green line shows the same model with additional sex interaction. P-values of model fit from likelihood ratio tests against model with just covariates. 59
Figure 4.5: P-value of sex interaction term across p-value thresholds. Sex interaction p-value from LR test. Black reference line at p=0.05, red line at p=0.001. Y axis presented on -10log scale. 60

Figure 4.6: AUC by Sex across p-value thresholds, linear scale, high-resolution. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI.....61
Figure 4.7: Comparison of AUC by Sex across p-value thresholds, Y axis shows p-value from de Long's test for uncorrelated ROC curves. Black reference line at p=0.05, red line at p=0.001. Y axis presented on -10log scale.....61
Figure 4.8: ROC curves by sex of PRS at p-value threshold 0.1012. P-value from de Long's test for uncorrelated ROC curves 63
Figure 4.9: PRS distribution by disease status and sex at p-value threshold 0.1012.....63
Figure 4.10: Quantiles by sex, for PRS at p-value threshold 0.1012 64
Figure 4.11: ROC curves of best thresholds by sex. Score 1= best Male score at pT 1x10-6, Score 2= best Female score at pT 1x10-5, p-values in each panel de Long's test of correlated ROC curves between score 1 and 2. 66
Figure 4.12: Quantile plot by sex for PRS score at pT $1x10^{-6}$.....67
Figure 4.13: AUC by Sex across p-value thresholds. Sensitivity - ambiguous SNPs included. 68
Figure 4.14: Sensitivity - MAF > 5% in target data 70
Figure 5.1: Venn diagram of variant overlap between input studies.....76
Figure 5.2: Manhattan plot of sex interaction p-values. Each dot represents a variant. Position is indicated along the x-axis, with the association of each variant given on the y-axis where higher -log10(p) indicates stronger association. Green dots represent SNPs used in the initial 15 SNP GRS (SPDL1 not included as MAF <1%). 81
Figure 5.3: QQ plot of meta-analysis, sex-interaction term.....81
Figure 5.4: PRS using sex-interaction term, model fit at selected low-resolution thresholds 82
Figure 5.5: AUC by sex across p-value thresholds for PRS based on the sex interaction effect size. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI.....84
Figure 5.6: PRS Score using Male SNPs (OR > 1 in sex-interaction term), effect sizes from standard GWA. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI. Linear x-axis (right); Log x-axis (left).....85
Figure 5.7: PRS Score using Female SNPs (OR < 1 in sex-interaction term), effect sizes from standard GWA. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI. Linear x-axis (right); Log x-axis (left).....86

Figure 5.8: Manhattan Plot of β_1 SNP term (for females) p-values. Each dot represents a variant. Position is indicated along the x-axis, with the association of each variant given on the y-axis where higher $-\log_{10}(p)$ indicates stronger association. Red horizontal line at $5x10^{-8}$ is the genome-wide significance line, blue line at $1x10^{-5}$ the genome-wide suggestive line. Green dots represent SNPs used in the initial 15 SNP GRS (SPDL1 not included as MAF <1%). Y axis truncated at 15 (MUC5B variant not shown). 88
Figure 5.9: QQ plot of meta-analysis, β_1 (SNP) term for females.....88
Figure 5.10: PRS association of β1 SNP term at selected pTs. R^2 on the liability scale assuming a population prevalence of 63 cases per 100,000.....91
Figure 5.11: AUC from PRS Score using β1 SNP term from sex interaction analysis. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI. Linear x-axis (right); Log x-axis (left). 91
Figure D.1: Deviance Residuals against fitted predictor for PRS at pT 0.1012.....118
Figure D.2: Leverage of individuals on PRS model fit at pT 0.1012.....118
Figure E.3: Region plot of rs6084435. Each point represents a variant with chromosomal position on the x axis and the -log(P value) on the y axis. Variants are colour coded by linkage disequilibrium with rs6084435. Gene locations are shown at the bottom of the plot 119

1 IDIOPATHIC PULMONARY FIBROSIS BACKGROUND

This section introduces the background for the project. It starts by summarising the medical background of idiopathic pulmonary fibrosis (IPF) and what is known about the genetics of IPF. The evidence for any sex differences within IPF and why these might exist is discussed. The aims of this project are set out in section 1.5 along with the outline for the rest of the project.

1.1 Idiopathic Pulmonary Fibrosis

Idiopathic Pulmonary Fibrosis (IPF) is defined by the American Thoracic Society and European Respiratory Society as:

"*A specific form of chronic, progressive, fibrosing interstitial pneumonia of unknown cause, occurring primarily in older adults, limited to the lungs, and associated with the histopathologic and/or radiologic pattern of usual interstitial pneumonia (UIP)*"[1]

It is a rare lung disease that results in irreversible scarring and progressively worsening lung function. There is no cure, and the prognosis is poor with an estimated median survival of 3-5 years after diagnosis[2-6].

1.1.1 Clinical Signs & Natural History

The main clinical symptoms are non-specific with a combination of a cough and difficulty breathing or dyspnoea the most common signs[1,7-9].

Lung function assessments typically show a reduced lung function with a reduced total

lung capacity, forced vital capacity (FVC) and a reduced capacity of the lung to diffuse carbon monoxide(DL_{co})[7].

The clinical course of the disease is unpredictable and may take one of several clinical forms. Most patients experience slow progressive decline in pulmonary function, until eventual death from respiratory failure. However, some progress much more rapidly whilst others remain stable with long survival[10].

Each year approximately 10% of patients experience acute exacerbations with a very high mortality rate[1,11].

1.1.2 Diagnosis

As IPF has no known inciting cause, diagnosing it requires both the exclusion of known causes of interstitial lung disease (ILD) and either a specific lung biopsy or high-resolution computed tomography consistent with UIP. Definitive diagnosis requires input from pulmonologists, radiologists, pathologists, and immunologists[1,8]. As accurate and timely diagnosis is challenging[7,12] the condition may be mismanaged in the early stages with patients receiving treatments such as immunosuppressive therapy meant for more common diseases, that do not benefit IPF patients.

1.1.3 Treatment

Management of IPF involves behavioural and lifestyle changes to reduce further injury to the lungs including support to stop smoking, and vaccines for diseases that can cause pneumonia (e.g., influenza and COVID-19). Supportive therapy includes supplemental oxygen and specific physiotherapy programs[8].

In 2015 two anti-fibrotic treatments with broad mechanisms of action were licensed which slow the disease progression by slowing decline in lung function– Nintedanib[13] and Pirfenidone[14].

1.1.4 Incidence and prevalence

Prevalence is defined as the proportion who are affected by a medical condition at any given time[15]. Incidence is a measure of the number of new occurrences of the medical condition within a specified time frame and is given as a rate[15], here cases per 100, 000 population per year (i.e., per 100,000 person years) has been used.

From section 1.1, IPF is considered a rare disease, but there are difficulties with determining the true incidence and prevalence. The diagnostic criteria (section 1.1.2) have only recently been established to try to standardise IPF definition and diagnosis. As a result, IPF case definitions for extracting IPF cases from electronic health record databases vary between studies, regions and over time[3,16].

The reported incidence amongst adults for cases of European descent (from Europe and North America) ranges from 0.22 – 17.43 per 100,000 person-years[3,16,17]. The reported prevalence ranges from 1.25 – 63 cases per 100,000 people[3,16,18]. Among the over 65s, the incidence and prevalence are higher, as expected, with a US based study reporting an incidence of 93.7 cases per 100,000 person-years and a prevalence of 494.5 cases per 100,000 people[19].

Overall, reported prevalence and incidence are higher in the US than in Europe and there is some evidence that IPF rates are increasing in all regions over time. This could

be due to increased diagnostic capabilities and recognition of the disease or due to aging populations[3,16].

1.1.5 Pathophysiology

The aetiology of IPF remains incompletely understood. The current working model is that repeated alveolar epithelial injury along with premature aging and senescence of the epithelial cells results in an inability to restore the epithelium and the lung fibrosis occurs as a pathological response[7]. In addition, host-defence mechanisms are thought to play an important part in the development of IPF[20]. The key genes that have been implicated so far in the development of IPF will be summarised in section 1.2.

1.1.6 Risk factors

Although the aetiology is unknown, there are several known risk factors that are associated with development of IPF.

Age: Incidence and prevalence increase with age and IPF primarily affects people from the 6th decade of life onwards[1,7,16,19].

Smoking: IPF is more common in people with a history of smoking, particularly heavy smokers with a smoking history of more than 20 pack-years[21,22].

Male sex: Much of the literature reports an increased prevalence of IPF amongst males[3,10,19,23]. See section 1.3.

Genetics: It has long been known that there is a genetic component to the development of IPF. The disease appears to cluster in families in a proportion of cases

(known as familial interstitial pneumonia) and a family history of ILD is a risk factor for IPF[24]. Within sporadic cases with no known family link, several genetic associations have been uncovered through genome-wide association studies (GWAS), see section 1.2.

Other risk factors with a possible association with IPF include gastroesophageal reflux, and chronic viral infections (e.g., Hepatitis C, Epstein Barr virus)[7,8]. Various types of dust exposure and agricultural work have been found to increase IPF risk in a recent meta-analysis[22] and there may be other, as yet unreported risk factors for IPF.

1.2 Genetics of IPF

1.2.1 Basic Genetic Concepts

Before considering the genetics of IPF, some genetic concepts that are central to the understanding of this project are covered.

DNA is made up of a sequence of base nucleotides: the purines, adenosine (A) and guanine (G); and the pyrimidines, cytosine (C) and thymine (T). These are arranged on long anti-parallel strands made up of 5 carbon sugars and phosphate groups in a double helix structure. Each rung is a base pair with A and T consisting of one pair, and C always pairing with G as shown in **Figure 1.1**. The left-hand strand is read from 5' to 3' (the carbon numbers counted clockwise from the oxygen 'O') and is also known as the +ve strand. The right-hand strand is read from 3' to 5' and is also referred to as the –ve strand. As the strands are complementary, each providing the template for the other, knowing the strand a specific base or sequence came from is vital to be able to match results between samples or studies.

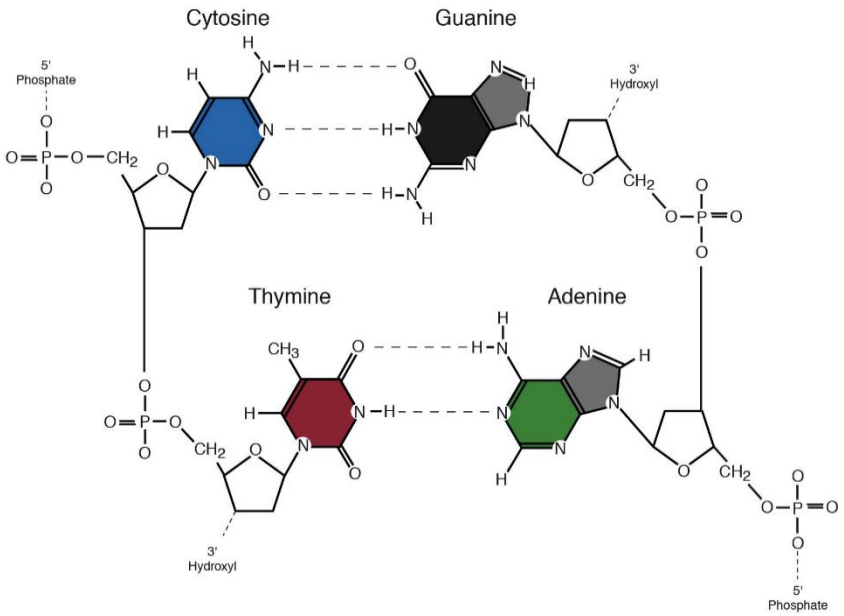


Figure 1.1: Molecular structure of DNA, showing the 2 base pairs[25]

A gene is a section of DNA that carries the genetic code for a specific protein or group of proteins. It is made up of several coding segments which contain the information to make the protein (exons), and non-coding segments (introns)[26]. For genes in the autosomal chromosomes, two copies are present, one per chromosome, each inherited from one parent.

During meiosis – the process in which the gametes (sperm and ovum) are formed – the parental chromosomes cross over and exchange DNA sections in what's known as recombination (**Figure 1.2**)[27].

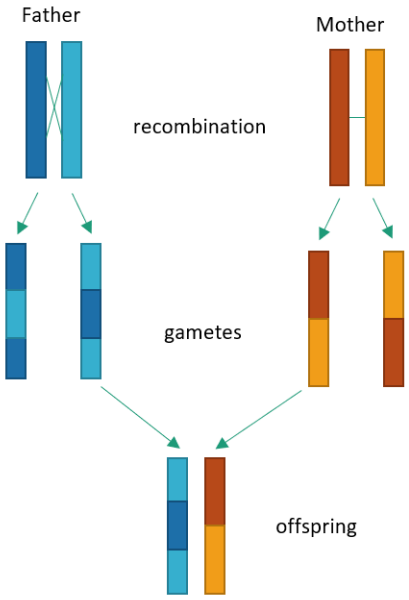


Figure 1.2: Chromosomal recombination during meiosis and inheritance from parents to offspring. Simplified example showing 1 chromosome

The complete DNA sequence is highly conserved, but variation exists. Variation that affects just one base pair at a specific point within the sequence is known as a single nucleotide polymorphism (SNP)[27]. The less common variant is known as the minor allele, and the proportion of these within the sample is the minor allele frequency (MAF).

Linkage disequilibrium (LD) is the correlation at a population level between (usually nearby) variants due to limited recombination. The set of variants that are inherited together is a Haplotype[27].

SNPs are common and are present across the genome in large numbers, allowing them to be used as markers for genetic variation[28]. SNP arrays directly genotype a set number of SNPs in the genome - with modern arrays containing 600,000 to several million SNPs[29]. Haplotype information from a reference panel can be used to impute

the genotypes of individuals for variants not in the array. This allows many more variants to be analysed than were originally genotyped including rare variants[30].

Structural genetic mutations – such as insertions, deletions, and copy number variants – also exist, but this project will focus specifically on SNPs as pre-existing genome-wide SNP data are available.

1.2.2 Overview of IPF Genetics

IPF is a complex binary trait with genetic and environmental risk factors (section 1.1.6). The disease itself is thought to occur on a continuum of genetic risk with familial IPF at one end and sporadic on the other (**Figure 1.3**). Within this project we'll be mainly focussing on unrelated individuals who are more likely to come from the right-hand side of this continuum.

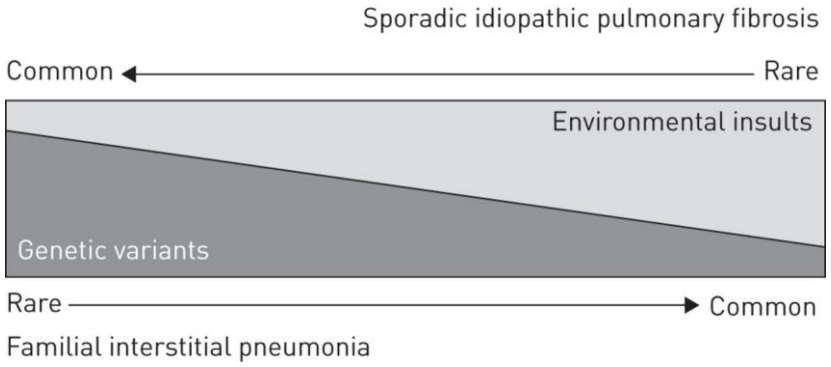


Figure 1.3: Familial to Sporadic continuum of IPF, from Kropski, J.A., T.S. Blackwell, and J.E. Loyd(2015)[31]

As the disease is rare, even someone with all known risk factors is unlikely to develop IPF. To date, several genetic associations have been identified through candidate gene studies and GWAS.

It helps to consider the liability threshold model of disease. Under this model we assume that individuals have an unmeasured continuous trait, the liability, which captures all genetic and environmental risk factors. An individual develops IPF when their total liability is higher than a specific critical threshold. The narrow sense heritability (h^2) is then proportion of variance on the liability scale that can be attributed to genetic variation[32]. To transform between the liability scale and the observed scale the population prevalence needs to be considered.

Figure 1.4 illustrates the genes implicated in IPF to date in the European population within a framework of effect size (OR) by allele frequency. Most genetic discoveries (across diseases) fall between the two dashed lines.

The white genes/variants in **Figure 1.4** are rare with large effect sizes and have been implicated in familial IPF, though they have been seen in some sporadic cases with no known familial link[31]. These mutations change the structure of the protein and are not are not identified by GWAS due to their rarity and as they are not generally SNPs. The variants indicated by black gene names are identifiable by SNPs and important in sporadic IPF. These are predominantly common variants with low to modest effect.

Compared to the other variants, the *SPDL1* risk variant has a lower frequency at 0.008. *SPDL1* is involved in mitosis and the risk variant is a SNP mutation that changes 1 amino acid in the protein from Arginine to Glycine. It did not reach genome-wide significance in GWAS studies[33] due to its low-risk allele frequency, however, the variant is present in the data that will be used in this thesis. *MUC5B* is somewhat of an outlier as a high effect common variant in an otherwise rare disease. It is found in the general (European) population at frequencies of around 11%[34], with 31-42% patients with IPF

carrying at least one risk allele[34,35]. Despite the large effect, the vast majority of people with a risk allele will not go on to develop IPF.

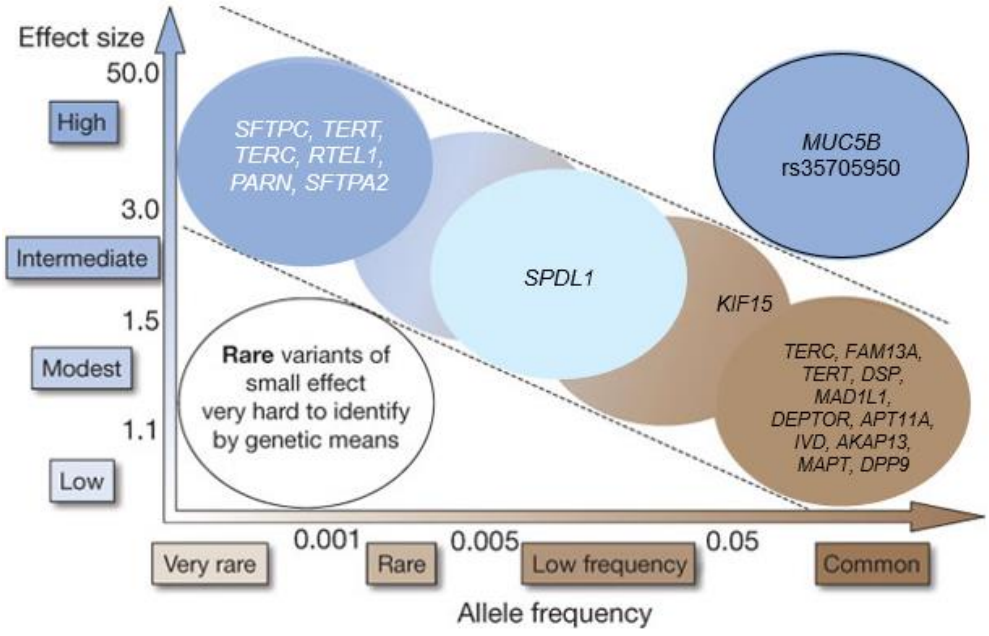


Figure 1.4: IPF genetic variants by risk allele frequency and strength of genetic effect (OR). Genes/variants in white text primarily involved with familial IPF, genes in black important in both sporadic and familial. Note that the effect size and frequency is illustrative. Adapted from TA Manolio et al. Nature 461, 747-753 (2009)[36]

No individual variants with low frequency and small effect size have been identified due to limitations in current genetic approaches, but these do contribute to the overall heritability of IPF.

The known genetic risk variants and the genes implicated from GWAS[34] and the SPDL1 variant from Dhindsa et al.[33] are detailed in **Table 1.1**. The effect sizes are from the largest IPF meta-GWAS to date[34] which combined data from 3 previous studies (UK[37], Colorado[38], Chicago[39]). Within these there are genes related to epithelial barrier function (*DSP, DPP9*). There are also genes related to cell division including telomere maintenance (*TERT, TERC*), cell cycle regulation (*KIF15, MAD1L1*).

Chrom	position	SNP	Gene	EA	EAF	effect size	p-value
3	44902386	rs78238620	*KIF15*	A	0.051	1.58	$5.12x10^{-10}$
3	169481271	rs12696304	*TERC*	G	0.278	1.31	$7.09x10^{-13}$
4	89885086	rs2013701	*FAM13A*	G	0.5	1.27	$3.30x10^{-13}$
5	1282414	rs7725218	*TERT*	G	0.666	1.39	$1.54x10^{-20}$
5	169015479	rs116483731	*SPDL1*	A	0.008	2.4	$7.55x10^{-07}$
6	7563232	rs2076295	*DSP*	G	0.464	1.46	$2.79x10^{-30}$
7	1909479	rs12699415	*MAD1L1*	A	0.415	1.28	$7.15x10^{-13}$
7	99630342	rs2897075	*7q22.1*	T	0.372	1.3	$3.10x10^{-14}$
8	120934126	rs28513081	*DEPTOR*	A	0.568	1.23	$1.20x10^{-09}$
11	1241221	rs35705950	*MUC5B*	T	0.138	4.84	$1.18x10^{-203}$
13	113534984	rs9577395	*APT11A*	C	0.797	1.3	$1.34x10^{-10}$
15	40720542	rs59424629	*IVD*	T	0.534	1.31	$7.30x10^{-16}$
15	86097216	rs62023891	*AKAP13*	A	0.303	1.27	$1.27x10^{-10}$
17	44214888	rs2077551	MAPT	T	0.813	1.41	$2.83x10^{-16}$
19	4717672	rs12610495	DPP9	G	0.306	1.31	$2.92x10^{-12}$

Table 1.1: Known variants associated with risk of IPF.

EA=Effect allele; EAF= Effect allele frequency. Effect size (odds ratios) and p-values from meta-GWAS[34]

It has been estimated that these known variants in **Table 1.1** (excluding *SPDL1*) account for 8 – 12% of the IPF risk in in the general European population, with the majority of that (5.9 – 9.4%) coming from *MUC5B*[40].

1.2.3 MUC5B

The most consistently reproduced genetic association in European cohorts with IPF lies in a SNP (rs35705950) within the promoter region of the Mucin 5B gene (*MUC5B*) on chromosome 11 with the T allele conferring risk[35]. It has been widely replicated and has been the dominant finding in all GWAS performed to date[37-39]. As it is found in the promoter region, the protein produced by the gene is not changed, rather, its expression is increased in lung tissue[35]. In addition, *MUC5B* is over-produced in patients with IPF, whether the risk variant is present or not – with *MUC5B* gene

expression more than 14 times higher in IPF patients compared to normal controls, irrespective of genotype[35] indicating that *MUC5B* expression is associated with IPF development.

The *MUC5B* variant does not appear to be associated with the development of interstitial lung disease secondary to other diseases such as sarcoidosis[41] and systemic sclerosis[41,42]. It has also not been implicated in other lung diseases such as Chronic Obstructive Pulmonary Disease (COPD)[43].

Prognosis appears to better for those who develop IPF and have the *MUC5B* variant[44]. This might suggest the existence of distinct genetic subtypes of IPF, with the *MUC5B* variant associated with a less severe form of IPF. However, it's also possible that the finding is due to index event bias[45].

1.3 Sex differences in IPF

This section considers the sex differences in IPF. That differences exist is undisputed. However, little is known about the underlying reasons and magnitude of effects.

Firstly, the use of the term 'sex' needs to be clarified. Biological sex is a designation assigned at birth based on chromosomes, gonads, hormones and genitals[46].

As this project deals predominantly with genetics, anyone whose reported sex did not match up with their chromosomal sex (as determined by the number of X chromosomes) was excluded from the analysis. As such, throughout this project, sex is used for subjects whose reported and genetic sex match.

1.3.1 Differences in IPF traits

Incidence and Prevalence

From section 1.1.6, the reported prevalence of IPF is higher in men. Estimating exact incidence and prevalence by sex is subject to methodological issues (section 1.1.4). However, the relative difference between the sexes in individuals that have contributed to recent studies is easier to extract. In the recent GWAS meta-analysis roughly 70% of the IPF cases included across 3 studies were male[34].

Prognosis/survival

Most observational studies report worse survival for males[4,23,47-51] that persists even after adjusting for smoking history. There is no difference in the proportion of patients presenting with a cough but coughing appears negatively associated with survival in men only[47]. There is also some evidence of more rapid deterioration in exercise tolerance over time compared with females[48].

1.3.2 Factors associated with sex difference in IPF risk

With IPF risk the best characterised of the sex differences within IPF, it is worth exploring the possible underlying reasons behind this.

Environmental Exposure

There may be differences in environmental exposures to as yet unknown risk factors. Historically more men smoked[52] and more men worked in agriculture, construction and other heavy industries. This means that men were exposed to relatively more inhaled irritants[53]. To give a related example – asbestosis and malignant mesothelioma are lung diseases with a strong link to inhaled asbestos. These have a large male

predominance with the incidence rate of mesothelioma in men up to 3.3 times that in women[54].

Diagnostic Bias

It is possible that historically differences in lifestyle factors like smoking accounted for a male predominance within IPF. Even with a true causal association, it is possible for such a factor to be removed from the population, without much noticeable difference in the case rates between sexes. Once male predominance is established, it becomes a core component of the pattern recognition used in the diagnosis and it becomes a self-perpetuating form of diagnostic sex bias. When clinicians are faced with otherwise identical case files, men are more likely to be given a diagnosis of IPF[55]. It is likely that a diagnostic bias exists and that men are over-diagnosed, and women under-diagnosed with IPF[55].

Physiological differences in Lung Function

There are physiological and physical sex differences in the lungs, where the sex hormones play a key part in development. The global lung function reference values model lung function throughout life separately for males and females[56] and in adulthood a number of key differences remain. Males have higher total lung volume and higher peak expiratory flow, but lower FEV1/FVC ratios and higher resistance. Female lungs are smaller, but more efficient[57]. A decline in lung function over time with decreasing lung elasticity and increasing connective tissue is part of the normal aging process in both sexes. In men this rate of decline is higher and starts earlier than in females[57]. The role of the sex hormones in development of IPF is not clear[58], but there is some evidence from in-vitro and animal studies that androgens exacerbate, whilst oestrogens are protective against fibroblast growth and subsequent fibrosis[59,60].

Genetic differences

Under the liability threshold model of disease development for a binary trait (section 1.2), a sex predominance can be explained through the existence of sex dependent liability thresholds, where, for females to develop the disease a higher liability threshold must be met – i.e., a greater number of risk alleles are needed to for a female compared to a male for the same increase in disease risk. The corresponding heritability is then expected to be higher for females compared to males[46].

It's also possible for gene-environment or gene-sex interactions to be present that affect the overall sex prevalence. A specific genetic variant may have differential effects by sex in either magnitude or direction of effect[61]. For example, a risk variant in males may be protective in females. Variants strongly associated with disease may not be identified in traditional GWAS if sex interactions are present, but not accounted for. Larger sample sizes are needed to account for the increase in complexity of the statistical model or sex-specific GWAS[62-64]. Bernabeu et al.[62] investigated sex differences across 530 traits in the UK Biobank and found evidence of gene-sex interactions in many traits, with modest increases in predictive accuracy when sex-specific effects were considered. There have also been findings in other lung diseases with a gene-sex interaction identified in COPD within the *HHIP* gene[63]. Within IPF there is some preliminary unpublished work to suggest that a gene-sex interaction may be present within the *DSP* variant.

Lastly, the sex chromosomes should be considered. Males possess just one copy of the X-chromosome whilst the female has two. Within females, the cells undergo a process known as X-chromosome inactivation to ensure that, generalyl, expression only comes from one allele[65,66]. However, it's not possible to elicit which X-chromosome is

expressed within any cell or tissue from genotype data alone. This poses several methodological issues when attempting to analyse this data[33]. As a result, most GWAS to date have focussed exclusively on the autosomal chromosomes[46].

In summary, the best characterised of the sex differences is the increased IPF risk in males. The underlying reasons are unclear and possible factors include environmental exposure, diagnostic differences, differences in underlying respiratory physiology between the sexes, or genetic differences.

1.4 Prediction models in IPF

Substantial differences exist within IPF in terms of the clinical course of the disease (section 1.1.1). Efforts to apply precision medicine and the creation of clinical prediction have focussed mainly on improving prognostic accuracy for subjects with an IPF diagnosis. A summary of selected models evaluated to date is shown in **Table 1.2.**

Score name	Variables included	Performance	Original Study
Prognostic			
CRP – clinical radiologic physiologic	age, smoking history, clubbing, extent of profusion of interstitial opacities, presence of pulmonary hypertension on the chest radiograph, % predicted TLC, PaO2 at the end of maximal exercise	Correlation with histopathological features as proxy for survival. Pearson's r = 0.37	2001, King et al.[67]
CPI - composite physiologic index of disease severity	extent of disease pn CT = DLco, FVC, FEV_1	Correlation with extent of disease on CT r^2=0.51.	2003, Wells et al.[68]
MRSS – mortality risk scoring system	age, history of respiratory hospitalisation, % predicted FVC, 24-week change in FVC	1 year mortality (C= 0.75 (0.71-0.79)	2011, du Bois et al.[69]
GAP – gender age physiology	Sex, Age, FVC, DLco	Mortality at 1,2,3 years C-index 69.1	2012, Ley et al.[70]
RISE - RIsk Stratification score	Medical Research Council Dyspnoea Score (MRCDS), 6-min walk distance (6MWD), CPI	3-year survival, AUC= 0.76	2012, Mura et al.[71]

MUC5B	Sex, FVC, DLco, *MUC5B* variant genotype	Survival time at 100-day intervals C=0.73(0.62-0.78)	2013, Peljto et al.[44]
PBMC gene expression + clinical markers	Expression of *CD28, ICOS, LCK*, *ITK* in PBMC cells; age, sex, FVC%	Transplant free survival AUC=0.79, SE=0.048	2014, Herazo-Maya et al.[72]
PI - Prognostic Indicator	118 gene expression panel in PBMC cells	Survival in High v low risk IPF cohort SHR 2.7; 95 % CI 1.9-3.9	2015, Huang et al.[73]
CALIPER (revised CPI)	extent of disease on CT using CALIPER computer algorithm = DLco, FVC, FEV_1	18-month Transplant free survival AUC=0.75	2020, Hosein et al.[74]
Diagnostic			
CDSS – clinical diagnostic scoring system	age, sex, smoking history, ethnicity, ILD family history, environmental exposures, connective tissue disorders, velcro crackles in lung	Differentiation of IPF from other ILDs, AUC=0.88	2021, Pastre et al.[75]
PI - Prognostic Indicator	118 gene expression panel in PBMC cells	IPF v healthy control AUC 0.96	2015, Huang et al.[73]

Table 1.2: Clinical prediction models in IPF

Due to differences in methods and outcomes measured, the performances of these scores are not readily comparable. In addition, small sample sizes used to create and validate these scores have meant poor reproducibility in the wider patient population and these prediction models have so far not been adopted into general use. So far, whilst sex has been included as an independent predictor, sex-specific differences in genomic biomarkers have not been considered.

As IPF is a rare disease, for clinical prediction scores to be used in screening to identify subsets of patients at high risk of developing IPF, the performance needs to increase substantially[76]. If there are genetic differences between the sexes, being able to incorporate these into risk prediction may offer advantages over existing models.

1.5 Objectives and Outline

The main aim of this project is to use polygenic risk scores to test whether there are differences in IPF risk prediction accuracy between males and females.

Within this broader question the focus is on 2 objectives.

1) To investigate the existence of any sex differences in the genetic architecture of IPF.

2) To test whether taking sex into account can help to improve prediction accuracy of a PRS.

Chapter 2 introduces the polygenic risk score and the main methods that will be used throughout the project to construct these scores and assess their performance whilst considering sex.

In Chapter 3 genetic risk scores incorporating the known IPF variants (**Table 1.1**) are assessed for both overall predictive accuracy and for sex differences. For this the non-sex-specific estimates from the largest IPF GWAS to date are used. These initial analyses function as an anchor point and comparison for further exploration.

Chapter 4 expands the analysis in Chapter 3 to create polygenic risk scores that can use all variants from the full GWAS summary statistics so that variants which did not reach genome-wide significance, but that do contribute to the heritability of IPF can be accounted for.

In Chapter 5 results from a sex-interaction GWAS are used. These are first meta-analysed to obtain effect sizes for the female specific SNP term and the sex interaction term. Unlike previous chapters scores are not just stratified by sex, but sex-specific effect estimates and sex interactions can be considered in the construction of the

scores. In this chapter an exploratory GWAS using the female specific estimates is also carried out to identify female specific associations.

Chapter 6 is a discussion of the overall findings, the strengths and limitations of this projects and suggests avenues for further research.

1.6 Summary

In this chapter the background to the project was introduced. IPF was described from both clinical and genetic perspectives. Sex differences and potential reasons for these were explored. This was followed by the current use of genetics and sex in the prediction and diagnosis of IPF. Lastly, we set out the overall objectives and outlined the further structure of this project.

2 POLYGENIC RISK SCORES BACKGROUND

This chapter introduces the polygenic risk score. This is the main statistical framework that will be used to approach the aims of this project. First the more general background is considered. Then methods for the assessment of performance of the scores are described. Lastly the applications of polygenic scores, more generally, and specific to this project are discussed.

2.1 Polygenic Risk Scores

Rather than focus on individual associations, the effect of multiple SNPs can be combined into a cumulative estimate of risk. Generally, the polygenic score provides a measure of individual risk based on the sum of risk alleles a subject possesses weighted by the effect size of each risk allele. For a binary trait like IPF the effect sizes are the beta coefficients (log Odds Ratios) taken from an initial training sample (base data) – a GWAS with as large a sample size as possible[77]. The score is then applied to a target sample with individual genotype and phenotype data as follows:

$$PRS_j = \sum_{i=1}^{N} \beta_i \times G_{ij} \qquad (2.1)$$

Where PRS_j is the score for the j^{th} person, β_i the effect size (log OR) of the i^{th} SNP, G_{ij} the number of risk alleles for SNP ij. N is the total number of SNPs in the score.

When considering the cumulative risk of known robust variants, the term genetic risk score (GRS) will be used in this project.

Individual risk variants identified from GWAS only go some way towards explaining the heritability of complex traits like IPF (section 1.2). Much of the phenotypic variation lies in the cumulative effect of many variants with modest effects that do not achieve statistical significance within GWAS[36].

Within this project the term polygenic risk score (PRS) is used when constructing scores that can consider all variants common to both the base and target data to take in the cumulative effect of all genetic risk. The ideal score includes all variants that contribute towards the development of IPF and none that do not. In the hypothetical scenario of perfect genetic effect size estimation from GWAS, the optimal PRS explains the variability in target sample equal to the total SNP heritability[78]. In practice the predictive power of the PRS is substantially lower. As not all SNPs will influence the risk of developing IPF, and there is uncertainty and sampling error in the estimation of individual SNP effect size, including the unadjusted effect sizes of all SNPs can produce poor scores if sample sizes are insufficient[77,79]. The other main issue is the LD in the genome which complicates the simple addition of aggregate effects[77].

2.2 PRS methods

All methods used to generate PRS aim to adjust for the issues detailed above by considering shrinkage of the effect size estimates and controlling for LD.

The classic PRS method is the clumping and thresholding method (C + T). SNPs in the target data are clumped to create a subset of partially independent SNPs and PRS scores are computed based on this subset of SNPs only. Multiple scores can be tried based on p-value thresholds (α) to optimise the score and get as close to the ideal PRS

score as possible (2.2). In this method, strength of association (SNP p-value) rather than effect size is prioritised when considering inclusion of variants.

$$PRS_j(\alpha) = \sum_{i=1}^{N} \beta_i(P_i < \alpha) \times G_{ij} \quad (2.2)$$

Clumping involves grouping SNPs based on their genomic position. P-values from the base data are matched to the target data for each SNP. Within each clump the top SNP is retained and any SNPs that are in LD with this top SNP are removed from the analysis. This creates a semi-independent set of SNPs and the PRS can then use the assumption of independent additive genetic effects.

As this project has applied aims the classical clumping and thresholding method of generating PRS will be used throughout.

2.2.1 PRSice

PRSice[80] is an open-source software package that uses R and PLINK to automate many of the required steps for the C+T method. Within PRSice the following steps are performed:

Ambiguous SNPs – which are SNPs where the variants are part of the same base pairing, but on opposite strands (i.e., C-G compared to G-C) – are excluded. This is done automatically to avoid strand mismatches (positive v negative sense) between the base and target data.

Clumping is applied with a default clump size of 250Kb and $R^2 > 0.1$, with values changeable.

Target data strand flipping is performed within the target data if needed to ensure that the alleles counted correspond to the effect size and direction from the base data.

Then polygenic scoring is performed to create a PRS at each specified p-value threshold. The fit of these scores is tested using a logistic regression model. Covariates can be specified. PRSice also calculates the R^2 – for binary data the default is Nagelkerke's R^2, but it is possible to request the liability R^2.

Several outputs are produced including a bar chart of the R^2 and association p-values of selected thresholds, a line plot of the score p-value across thresholds, and a quantile plot. Also available are text files of the summary results and raw results for all thresholds. As the interest lies in the sex-stratified results, this last file is the most useful, containing subject level results and one column per p-value threshold.

2.3 Assessing performance and Predictive accuracy

This section focusses on the assessments of performance and predictive accuracy of polygenic scores for binary traits under the liability threshold model using case-control studies. As the proportion of cases is far higher than found in the general population evaluations of performance are affected by ascertainment and enrichment of cases.

2.3.1 Distribution of Score

The PRS distribution should approximate the normal distribution[77]. This can be assessed with descriptive statistics and kernel density plots with non-normality indicating violations of the underlying assumptions – for example the inclusion of many strongly correlated SNPs or population stratification issues.

2.3.2 Association

Testing the strength of association between the computed PRS score and phenotype within the target data is one of the key score optimisation measures. For a binary outcome, this can be done using a generalised linear model with a binomial distribution $\sim B(p, n)$ and mean μ using a logit link

$$\eta = g(\mu) = \log\left(\frac{\mu}{n-\mu}\right) = \log\left(\frac{p}{1-p}\right) \tag{2.3}$$

, i.e., logistic regression.

The linear predictor η models the log of odds for a subject with a given score and covariate pattern to be an IPF case (equation (*2.4*)).

$$\eta_i = \beta_0 + \beta_1 PRS\ Score_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \cdots + \varepsilon_i \tag{2.4}$$

Here η_i is the log of odds for individual i, with $PRS\ Score_i$ and covariates X_{ji}. β_0 is the intercept and ε_i is an error term with normal distribution $\sim N(0, \sigma^2)$. β_j the difference in log of odds for a 1 unit increase in the j^{th} covariate, i.e., the log odds ratio. Here e^{β_1} is the odds ratio for a 1 unit increase in PRS score. A score not associated with phenotype corresponds to a $\beta_1$of 0.

The p-value of the PRS score is obtained from a likelihood ratio test with 1 degree of freedom of the null model (containing any covariates but excluding the score) against the full model including score with a null hypothesis of β_1 = 0. As many p-value thresholds are considered, multiple testing needs to be accounted for. Each successive PRS score is incremental with a relatively small number of additional SNPs. As such, the

tests are not independent and a correction for multiplicity that assumes independence will be excessively conservative. From a preliminary empirical investigation a significance threshold of $p < 0.001$ has been suggested[80]. The logistic regression allows further model and goodness of fit investigation using standard methods such as deviance or AIC to compare relative model fit and examination of residuals, influence and leverage for subjects that may be driving the conclusions.

2.3.3 Discriminative ability and AUC

One of the most intuitive assessments of the PRS performance is to consider the ability of the score to discriminate between cases and controls. This can be done with the area under the receiver operator characteristics curve (AUC)[77]. It considers the performance of the PRS at the full range of sensitivities and specificities and distils that into a single number. The sensitivity is defined as the proportion of people with the disease who are correctly identified as diseased and the specificity is the proportion of non-diseased people who are correctly identified as not diseased at the chosen cut-off[81].

Consider a sample with *m* cases and *n* controls, with $X_i, \dots, X_m$ the values of score for cases, and $Y_j, \dots, Y_n$ the values of score for controls. The empirical sensitivity and specificity at each score (cut-off point) are given by:

$$sensitivity(score) = \frac{1}{m}\sum_{i=1}^{m} 1(X_i \geq score) \tag{2.5}$$

$$specificity(score) = \frac{1}{n}\sum_{j=1}^{n} 1(Y_j < score) \tag{2.6}$$

The empirical ROC curve is produced by calculating the sensitivity and specificity at all calculated values of score and plotting sensitivity(score) against [1 - specificity(score)], with the AUC given by

$$AUC = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(X_i, Y_j), \qquad \psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases} \tag{2.7}$$

The area under the ROC curve is equal to the Man-Whitney U-statistic. This is an unbiased estimate of the probability that a random control will have a score less than or equal to a random case.

The AUC can be interpreted as the probability that a case -control pair of subjects is classified correctly by the PRS. The larger the AUC the better the discriminative ability of the PRS with an AUC of 1 indicating perfect discrimination and 0.5 not better than chance. A key property is that the AUC is independent of the proportions of cases and controls being classified and is therefore not affected by ascertainment in case control studies [76]. However, there is a maximum value of the AUC for any specific disease that depends on the underlying heritability and disease prevalence which is independent of the amount of genetic risk explained by the PRS[76,82].

AUC can be estimated for males and females separately and their respective ROC curves can be constructed and compared. Comparison can be done visually, or by comparison of performance at specific thresholds. AUCs from two PRS scores can also

be compared with a formal hypothesis test such as DeLong's test for correlated ROC curves[83] using the properties of the generalised Mann-Whitney U statistic.

A z-score can be calculated as shown below, using the AUC of each ROC curve ($\hat{\theta}^A$ and $\hat{\theta}^B$) along with their variances and the covariance between them (*2.8*).

$$z = \frac{\hat{\theta}^A - \hat{\theta}^B}{\sqrt{Var\left[\hat{\theta}^A - \hat{\theta}^B\right]}} = \frac{\hat{\theta}^A - \hat{\theta}^B}{\sqrt{Var\left[\hat{\theta}^A\right] + Var\left[\hat{\theta}^B\right] - 2Cov\left[\hat{\theta}^A, \hat{\theta}^B\right]}} \quad (2.8)$$

The Wald method can be used for the 95% Confidence interval of the AUC (*2.9*).

$$AUC \pm 1.96 \times \sqrt{Var(AUC)} \quad (2.9)$$

For details of the calculation of the variance and covariance see APPENDIX A.

For comparison of AUCs from a stratified score the ROC curves are uncorrelated, and the p-value can be calculated using an unpaired t-test with unequal sample size and variance.

2.3.4 Variance explained

For a quantitative trait the proportion of variance explained or R^2 is intuitive and directly estimable from the linear regression. For a binary trait only pseudo-R^2 measures are available.

Nagelkerke's R^2 does not account for ascertainment and the estimates are biased on the liability scale where the prevalence and case/control ratio are not equal[77]. To

adjust for ascertainment, a Liability R^2 accounting for enrichment of cases has been proposed[84] and used[40]. It is defined as:

$$Liability\ R^2{}_{cc} = \frac{R^2_{0_{cc}} C}{1 + R^2_{0_{cc}} C\theta'} \tag{2.10}$$

Where $R^2_{0_{cc}}$ is the proportion of total variance explained by the PRS score on the observed probability scale in a case-control study which comes from the coefficient of determination on the observed scale from a linear regression, corrected for sampling proportion as given by:

$$R^2_{0_{cc}} = \frac{Var(fitted\ values)}{P(1-P)} \tag{2.11}$$

and where C and θ are given as follows:

$$C = \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)} \tag{2.12}$$

$$\theta = m\frac{P-K}{1-K}\left(m\frac{P-K}{1-K} - t\right), \tag{2.13}$$

Here P is the proportion of cases in the study and K the population prevalence, m is the mean liability for cases and t is the liability threshold (above which disease occurs and which is determined by the population prevalence) with z the normal density at t.

As for the p-value of the association, the R^2 of the PRS score is the incremental R^2, i.e., the increase in R^2 due to the addition of PRS to the model. Low values reflect the effect of sampling variation on the variance that is explained by a score[82].

2.3.5 Risk Classification

To examine the increase in disease risk in those with the highest PRS scores compared to the lowest the sample is divided into a number of quantiles based on PRS score, for example 10 equally sized deciles. These quantiles are then used in place of the continuous score in logistic regression. Covariates can be adjusted for and the quantile split can be stratified by key categorical parameters such as sex. As the score is categorised it does not assume linearity across quantiles, instead, with enrichment of cases, a plot of the quantiles against OR is expected to show an exponential distribution with a much larger relative increase in OR for higher quantiles[77].

2.3.6 Validation

The above measures can be used to optimise the PRS for performance in the target data. As the target data are a training sample, this is not enough to determine PRS score validity. In order to validate the findings the above measures should be repeated for the optimised score in a separate dataset for out of sample prediction[85].

2.4 Applications of PRS scores

Interest in PRS scores has increased in recent years and scores have been developed for multiple diseases and traits. An online catalogue of published polygenic scores, the

PGS catalog, has been established to promote sharing, reproducibility and traceability[86]. Applications include scores to assess the genetic overlap between traits or diseases, assessment of missing heritability, the identification of high-risk subgroups within a population or disease and individual risk prediction. They can also inform stratified medicine approaches through PRS scores to identify those at high risk of worse disease or identify subclasses of disease that respond better to specific targeted therapies[87]. PRS for identifying high risk subgroups in common diseases such as cardiovascular disease and breast cancer have been developed[88]. PRS scores can be used in the investigation of sex differences in genetic architecture of traits and recently PRS scores have been used at a large scale to detect sex differences in common traits, both binary and continuous, within the UK Biobank[62].

For individual risk prediction, there is some way to go before PRS scores reach clinical utility[89]. So far PRS scores have been largely developed in people of European ancestry and these do not generalise well to other populations[90].

2.5 Applications in IPF and this project

Within IPF there is one published PRS. Allen et al. used a PRS approach to investigate the polygenicity of IPF and found that a PRS excluding all known risk loci was significantly associated with IPF phenotype and estimated to account for around 2% of phenotypic variation[34] and thus indicates that there are genetic variants that contribute to IPF risk that have not yet been discovered through GWAS, either because of underpowered GWAS or the cumulative effect of many variants with individually small effects. In additional work (PhD thesis), Allen used PRS to investigate the overlap

between IPF and other lung traits and found some overlap between lung airflow traits and IPF risk[91].

Usually, PRS scores are optimised for their overall association with the target phenotype, by identifying the p-value threshold with the strongest association in the target data. The principles described can instead be used to optimise the PRS for a difference in prediction accuracy between the sexes. The distribution of any sex differences can be characterised by considering the range of p-value thresholds that produce sex differences in the performance of the PRS. This will help with the first of the aims outlined in section 1.5.

For the second aim, PRS scores will be optimised for discriminative ability within each sex to identify if sex-specific scores can improve predictive accuracy.
For effect estimates, first the non-sex-specific estimates from GWAS will be considered (Chapters 3 and 4). With a male predominance in the sample these effect estimates may apply less well to females if a sex difference in genetic architecture exists. In Chapter 5 sex-interaction and sex-specific effect estimates will be explored.

2.6 Summary

In this chapter the PRS for summing cumulative genetic risks was introduced with a focus on the clumping and thresholding method. The methods for assessment of PRS performance in a binary trait were described.
Common applications were briefly described before focussing on how PRS scores have been implemented in IPF and how they could help to explore the objectives of this project. Next the data will be introduced, and the cumulative effect of the known risk variants will be investigated.

3 GENETIC RISK SCORE WITH KNOWN IPF RISK VARIANTS

This chapter introduces the studies and resulting datasets that will be used in this project. The effect size estimates come from a non-sex specific meta-GWAS as this was the most powerful analysis of IPF risk to date and data were available. The male predominance in IPF means that even using effect estimates that are not sex-specific will be able to show differences in performance between male and female PRS scores if a difference in genetic architecture exists between the sexes.

The sex-agnostic data are used to create a GRS of the 15 known IPF genetic variants that were introduced in section 1.2 using R version 4.0.1[92]. This first analysis therefore considers just the variants identified in the most powerful analyses to date that together make up a large proportion (8-12%) of the IPF risk in the general population[40]. This score is stratified by sex, explored, and assessed for performance and sex difference.

As PRSice will be used to perform the clumping and thresholding over a range of p-values, a technical replication is performed in the last part of this chapter using just the 15 GRS variants.

3.1 Methods

3.1.1 Data

As stated in chapter 2, the base data consists of GWAS summary statistics. For this section these statistics were from the largest IPF GWAS to date[34] and are the results of a meta-analysis of 3 studies with cases and controls of European ancestry[37-39], GWAS

catalog GCST009758. IPF is a rare disease, and the sample size is still small in GWAS terms with 2668 cases and 8591 controls. Within the cases 1811 are male, 806 female and 51 of unknown sex.

For the GRS creation and analysis only the 15 variants known to be associated with IPF from section 1.2.2 were included.

The target data, used to test the performance of the GRS is comprised of cases of European ancestry from the United States, United Kingdom and Spain, with controls from the UK Biobank[93]. The controls were chosen to match the age, sex and smoking history distribution found within the cases. It will be referred to as the UUS data going forward. Subjects included in this dataset have passed several QC steps. Full details of the target data QC procedures can be found in the Allen et al. data supplement[34].

Subjects were excluded if they had an individual call rate < 95%, if biological sex mismatched the genetic sex, or if they exhibited high heterozygosity with a chosen cut-off of > 5 standard deviations above the mean. Subject data were checked to identify any that had been mis-classified as IPF cases. Only subjects with genetically determined European ancestry were included and subjects were excluded if they were found to be related to other cases in the UUS sample or base data set GWAS. These steps resulted in a final sample of 793 unrelated IPF cases.

Controls were selected from the UK Biobank to include only unrelated individuals of European descent without interstitial lung disease who were not already controls included in the base data. Of those meeting these criteria, 10 000 were selected to follow a similar age, sex and smoking history distribution as had been found in the

cases. One of the original controls later withdrew consent, so analyses are based on 9999 controls.

For the data itself, information was present on the disease status, sex, first 10 principal components. Genotype information on known genetic risk variants was present in dosage form, corresponding to the number of risk alleles present. Some variants - rs12696304, rs35705950 and rs59424629 – were genotyped directly, with the dosage data therefore containing discrete 0/1/2 values. Others were imputed giving a continuous value between 0 and 2. Imputation was performed using the Michigan Imputation Server[94] using the Haplotype Reference Consortium as the reference panel. Only variants in the reference panel that were in Hardy-Weinberg equilibrium ($P > 10^{-6}$) with a call rate of > 95% and MAF > 1% were considered for imputation. Rarer variants (with MAF < 1% in the current study) are present in the data where these were genotyped directly or where the MAF in the study was less than found in the reference panel. Within the 15 known risk variants, the lowest imputation R^2 was 0.79 for rs2077551, and considered well imputed.

3.1.2 GRS calculation using R

A GRS was calculated using the SNPs and effect sizes of the 15 known genetic risk variants for IPF as found in **Table 1.1**. A score was derived for each person in the UUS target data as follows:

$$GRS_j = \sum_{i=1}^{15} \beta_i \times G_{ij} \quad (3.1)$$

Where GRS_j is the score for the j^{th} person, β_i the effect size (log OR) of the i^{th} SNP, G_{ij} the number of risk alleles for each SNP *ij*. R code for the calculation of the genetic risk score using has been included in APPENDIX B.

3.1.3 GRS Association with phenotype

The GRS was tested for association using a logistic regression model, with disease status as the outcome and score as the predictor.

Principal components were added to adjust for population structure with sex added as the primary covariate.

$$IPF_i = \beta_0 + \beta_1 Score_i + \beta_2 Sex_i + \beta_3 PC_{1i} + \cdots + \beta_{12} PC_{10i} + \varepsilon_i \quad (3.2)$$

Where IPF_i is the log of odds of having IPF for subject i with $Score_i$, Sex_i and PC_{1i} to PC_{10i}, where $PC_1$to PC_{10} are the genetic principal components. β_j is the log(OR) for a 1 unit increase in the j^{th} covariate.

To check the effect of Sex, a score by sex interaction was added, producing the following model.

$$IPF_i = \beta_0 + \beta_1 Score_i + \beta_2 Sex_i + \beta_3 Score_i Sex_i + \beta_4 PC_{1i} + \cdots + \beta_{13} PC_{10i} + \varepsilon_i \quad (3.3)$$

Within the analysis, the sex coding was female=0 and male=1. Therefore, the log(OR) for a 1 unit increase in Score for a female is given by β_1, with the log(OR) for a male given by $\beta_1 + \beta_2 + \beta_3$.

As (*3.2*) and (*3.3*) are nested, a likelihood ratio test with 1df was used to determine if model fit was improved by the addition of the sex interaction term.

3.1.4 GRS Discriminative ability

Area under the curve (AUC) for the ROC curve of the score was calculated empirically using the *pROC* R package[95]. First the overall AUC was calculated, and this was then stratified by sex by splitting the UUS sample into male and female samples and creating the ROC curve and calculating the AUC for each sample separately. The score cut-off point and performance at Youden's index were also calculated. The two sex-specific ROC curves were compared using an unpaired t-test with unequal sample size and variance (section 2.3.3).

3.1.5 GRS Clinical utility

Score thresholds can be generalised to the population by considering their positive predictive value (PPV) and negative predictive value (NPV). The PPV is defined as the probability that someone with a positive result actually has the outcome of interest, while the NPV is the probability that someone with a negative result does not have the outcome of interest[81].

GRS score was split into quantiles and the PPV was calculated at the cut-off points between the quantiles and at the Youden's index using previous prevalence estimates described in section 1.1.4, for the overall score and by sex if indicated.

$$PPV = \frac{Se \times Pr}{Se \times Pr + (1 - Sp) \times (1 - Pr)} \quad (3.4)$$

With 95% confidence intervals calculated using the Wald method [96].

$$PPV \pm 1.96 \times \sqrt{Var(PPV)} \tag{3.5}$$

Where the variance is given by

$$Var(PPV) = \frac{[Pr(1-Sp)(1-\mathrm{Pr})]^2 \times \frac{Se(1-Se)}{n_1} + [PrSe(1-Pr)]^2 \times \frac{Sp(1-Sp)}{n_0}}{[SePr + (1-Sp)(1-Pr)]^4} \tag{3.6}$$

Here Se is the sensitivity, Sp the specificity, Pr the prevalence, and n_1 and n_0 the number of subjects with and without disease respectively.

Firstly, prevalence estimates from the general population which range from 1.25 to 63 cases per 100,000 people[3,16,18] are used. Secondly, we considered those > 65 years of age, where the prevalence estimate is 495 per 100,000[19].

3.1.6 Sensitivity – Effect of individual variants on GRS

To check if any one variant was driving the performance of the GRS and sex difference, a sensitivity analysis was performed where the AUC, of GRSs with each variant removed in turn, was calculated by sex and plotted.

3.1.7 Sensitivity – Using PRSice to create GRS

The model fitted in PRSice was the same as was used on the GRS of the 15 SNPs *(3.2)*.

PRSice automatically removes ambiguous SNPs (Section 2.2.1). The base and target data were both genotyped using the same human genome reference build and strand direction of the genotyping for both base and target data are known in this study. It is therefore possible to match the ambiguous SNPs directly with their effect size and use them to create a score to determine the effect of excluding these ambiguous SNP. A copy of the PRSice source code was edited so that the text file used to exclude the ambiguous SNPs always remains blank, thereby retaining all SNPs at this step. No other changes were made to the code.

The following GRS new scores were produced and compared:

- PRSice standard GRS excluding ambiguous SNPs
- R GRS (methods in previous chapter) excluding ambiguous SNPs
- PRSice GRS including ambiguous SNPs

3.2 Results

3.2.1 GRS calculation using R

Consistent with other IPF cohorts, the proportion of male cases in the UUS data are much higher than female cases (75.3%) (**Table 3.1**).

Sex	Disease status		**Total**
	Control	Case	
Female	2790 (27.9%)	196 (24.7%)	2986 (27.67%)
Male	7209 (72.1%)	597 (75.3%)	7806 (72.33%)
Total	9999 (92.65%)	793 (7.35%)	10792

Table 3.1: UUS data case and control by sex

The cases have a higher score on average, with the mean score in cases (4.935) above the 75th percentile of the controls (4.301) (**Table 3.2**) but there is a large amount of overlap between the distributions with a comparable range (**Figure 3.1**).

Statistic	Control (N=9999)	Case (N=793)
Mean (SD)	3.758 (0.957)	4.935 (1.149)
Median	3.630	4.925
IQR	3.078 – 4.301	4.126 – 5.708
Min, Max	1.108, 8.271	1.945, 8.703

Table 3.2: Overall summary statistics of the Genetic Risk Score

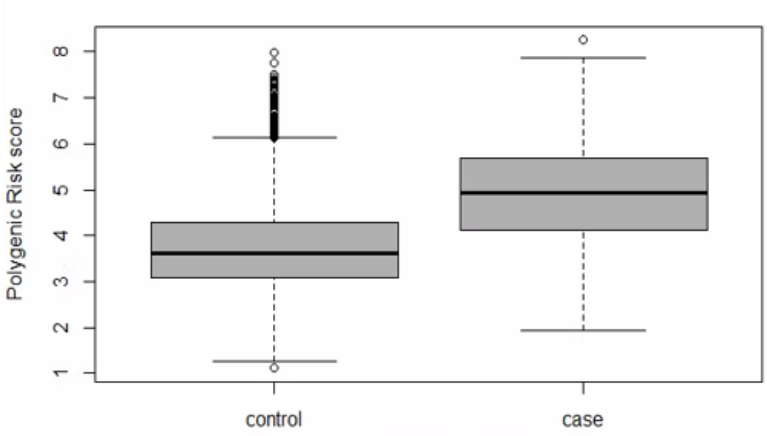


Figure 3.1: Boxplot of distribution of genetic risk score by disease status. Box represents the main body of the data, bound by the 25th and 75th percentile; outliers shown by open circles.

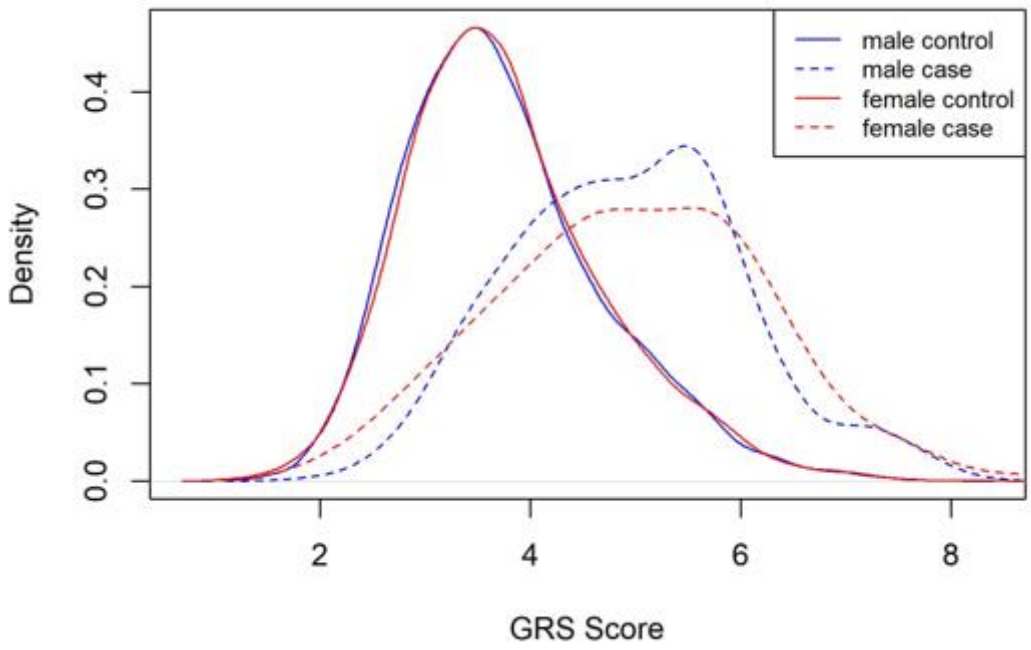


Figure 3.2: Genetic Risk Score distribution by disease status and sex for 15 variants with known IPF risk

Plotting the densities by status and sex (**Figure 3.2**) further shows the distribution of the GRS. There is a clear difference between the distributions of controls and cases, but little difference between the sexes. No major deviations from normality are observed in any of the groups.

3.2.2 GRS Association with phenotype

Summary results of the logistic regression association testing are presented in **Table 3.3**. From this, the genetic risk score constructed using the meta-analysis GWAS effect sizes of the 15 susceptibility variants is associated with disease status with a higher score corresponding to higher risk of disease.

Adding principal components to the model did not change the interpretation of the score coefficient, with a 1 unit increase in the GRS score corresponding to an increased OR for developing IPF of 2.67. The AIC of the model containing score and the 10

principal components did decrease despite the increase in complexity of the model. Overall, it suggests that there is some effect of the population structure when considering model fit, but that population stratification within this study is not likely to be a large issue in terms of the overall GRS.

Model	Score effect size (95% CI)	Score OR (95%CI)	p-value	AIC
Score	0.982 (0.913, 1.051)	2.67 (2.49, 2.86)	$1.32x10^{-170}$	4813.3
Score + 10PCs	0.981 (0.910, 1.053)	2.67 (2.49, 2.87)	$1.66x10^{-160}$	4557.1
Score + sex + 10PCs	0.982 (0.911, 1.054)	2.67 (2.49, 2.87)	$1.49x10^{-160}$	4552.1

Table 3.3: Logistic regression of GRS against disease status model summary results

Adding sex to the model, further improved the fit of the model (**Table 3.3**). The individual parameter estimate was 0.244 (CI: 0.063, 0.431; p 0.009). Exponentiation gives us an OR of 1.28 (CI: 1.06, 1.54) for development of IPF in the male sex compared to the female sex. However, this is a case control study where there was an attempt to control for sex distribution within the controls. This significant result indicates that control was not completely achieved, with a higher proportion of males in the cases compared to the controls (**Table 3.1**). As our primary aim concerns the investigation of sex differences, sex will be added to all models, but the parameter should not be interpreted in isolation.

The score by sex interaction term (equation *(3.3)*) did not significantly improve the model (Likelihood ratio p=0.6943). This indicates that the effect of score on the phenotype is not different between the sexes.

3.2.3 GRS Discriminative ability

The AUC for the male score is slightly higher than for the females score (**Table 3.4**). Visually the ROC curves are very similar, though at specificities below 0.6 the female ROC curve appears to have slightly worse performance, with corresponding sensitivity lower compared to the male and overall curves (**Figure 3.3**).

		Youden's index		
	AUC (95%CI)	threshold score	sensitivity	specificity
Overall	0.787 (0.768, 0.806)	4.146	0.747	0.705
Male	0.791 (0.769, 0.814)	4.146	0.747	0.707
Female	0.773 (0.733, 0.812)	4.180	0.745	0.711

Table 3.4: Predictive accuracy of GRS using Receiver Operator Characteristics

Comparison of male and female ROC curve using De Long's test for two ROC curves gave no evidence of a difference in performance between the curves at p=0.3845.

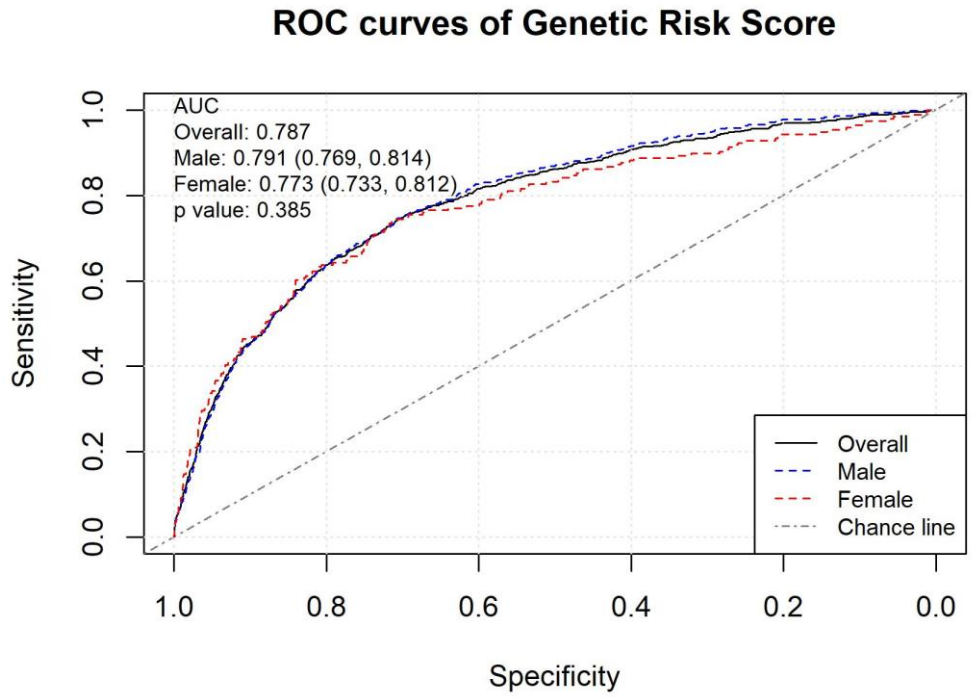


Figure 3.3: ROC curves of Genetic Risk Score, overall and by sex

3.2.4 Clinical utility of GRS score

The PPV of the overall score, at quintiles of GRS score and the Youden's index, for selected IPF population prevalence estimates is shown in **Table 3.5**. For illustration, assuming a GRS score of 4.146 the prevalence of 1.25 - 63 per 100 000 gives a PPV of 0.00317% to 0.159%, or of 100 000 people with a risk score above the threshold, 3 - 159 will have/develop IPF.

GRS Score	%[1]			PPV% (95% CI) at Specific Prevalence (x100 000)		
		Sens	Spec	1.25	63	495
2.99289	20%	0.965	0.213	0.00153	0.0772 (0.0772, 0.0773)	0.6062 (0.6061, 0.6063)
3.4666	40%	0.895	0.423	0.00194	0.0978 (0.0978, 0.0978)	0.7666 (0.7664, 0.7668)
3.9351	60%	0.792	0.631	0.00268	0.1352 (0.1351, 0.1352)	1.0568 (1.0564, 1.0573)
4.146		0.747	0.705	0.00317	0.1595 (0.1594,0.1596)	1.2448 (1.2442, 1.2454)
4.659	80%	0.588	0.831	0.00434	0.2184 (0.2183, 0.2186)	1.6982 (1.6970, 1.6994)

Table 3.5: PPV at specific GRS score thresholds and population prevalence estimates; [1] Percentile of score.

Taking the estimated prevalence in those > 65 years, improves the PPV markedly compared to the full population estimate (**Table 3.5** and **Figure 3.4**), but even at a cut-off at 80% (just the top 20% of scores above the threshold), of every 1000 people with a score of 4.659 or above, 17 will have/develop IPF

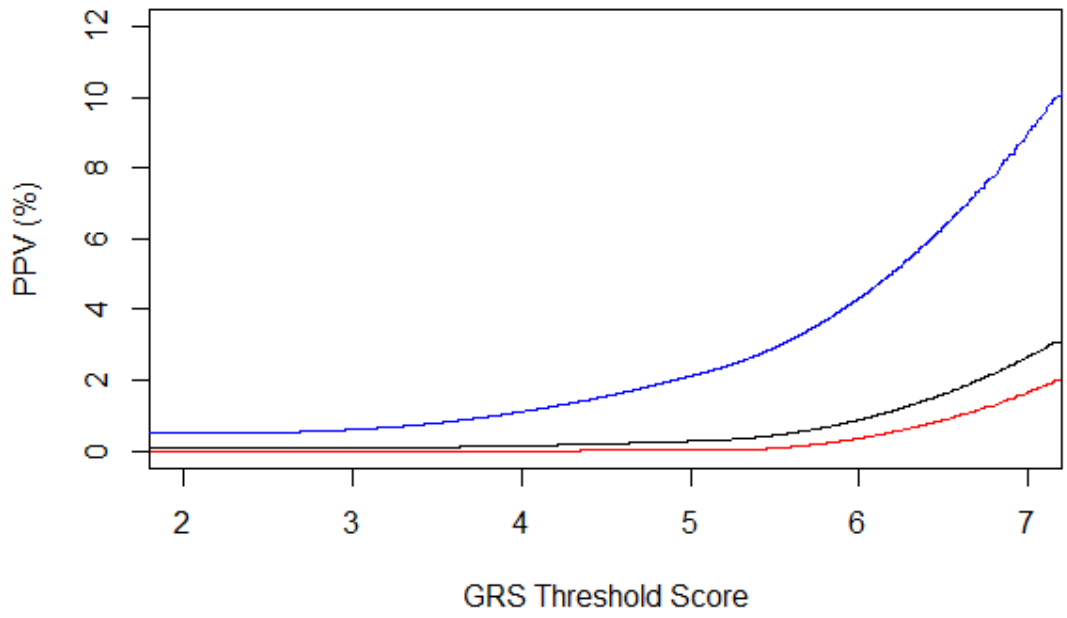


Figure 3.4: PPV of the 15 SNP GRS score at selected prevalence. Red line = 1.25, black= 63 and blue=495 cases per 100,000

3.2.5 Sensitivity – Effect of individual variants on GRS

When excluding each variant from the GRS in turn and plotting the results (**Figure 3.5**) the rs35705950 variant in *MUC5B* has the largest effect on the GRS by far. Despite this, the GRS score excluding *MUC5B* is still significantly associated with IPF risk and the AUC is greater than 0.5 for both sexes, indicating a discriminative ability of the score better than chance. Any other individual variant only has a small effect on the overall GRS. As for the 15 SNP GRS, the female point estimate of the AUC remains lower across all exclusions, though not significantly so.

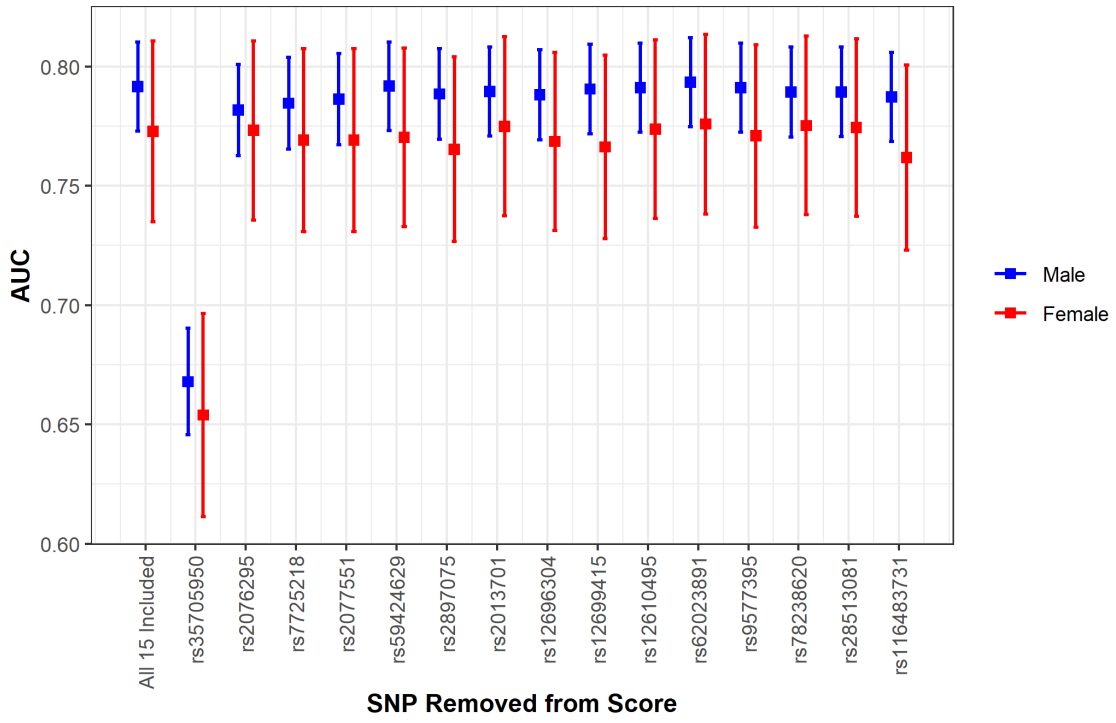


Figure 3.5: Effect of removing individual variants on performance of GRS

3.2.6 Sensitivity - using PRSice to generate GRS

When PRSice v1.25 was used to generate a GRS for the 15 known variants only 12 of the 15 SNPs were included in the score. The 3 ambiguous SNPs that were removed by PRSice before risk scoring are presented in **Table 3.6.**

chromosome:position	rsid	Gene	EA	non-EA	OR
3:44902386	rs78238620	*KIF15*	A	T	1.58
3:169481271	rs12696304	*TERC*	G	C	1.31
13:113534984	rs9577395	*APT11A*	C	G	1.30?

Table 3.6: Ambiguous SNPs removed from GRS by PRSice

The 12 SNP score produced by PRSice was highly associated with IPF status and had a p-value of 9.11×10^{-156} (**Table 3.7**). This is less significant than that obtained originally

using the dosage data and the manual calculation across 15 SNPs where the same model gave a p-value of 1.49×10^{-160}.

When the GRS score was recalculated manually in R with just the 12 non-ambiguous SNPS, the score p value was 3.82×10^{-156} – incredibly close to the PRSice generated score. From **Table 3.7** the performance is similar too. Agreement of performance in males was very high. There is a bit more of a difference between the female performance (0.778 PRSice compared to 0.770 with R) but the confidence intervals overlap and using de Long's test to compare the 2 scores for females (R v PRSice), did not show a significant difference (p=0.144).

			association	ROC AUC		
Method	N SNPs	Genotype	Score pval[1]	male (95% CI)	female (95%CI)	Pval[2]
PRSice	12	Hard Call	9.11×10^{-156}	0.783 (0.761, 0.805)	0.778 (0.739, 0.817)	0.796
R	12	Dosage	3.82×10^{-156}	0.785 (0.763, 0.808)	0.770 (0.730, 0.809)	0.469
PRSice[#]	15	Hard Call	2.72×10^{-160}	0.789 (0.767, 0.811)	0.782 (0.743, 0.821)	0.723
R	15	Dosage	1.49×10^{-160}	0.791 (0.769, 0.814)	0.773 (0.733, 0.812)	0.385

Table 3.7: PRSice v1.25 compared to manual code results.

[#]PRSice modified to allow for inclusion of ambiguous SNPs

[1]p-value of score from model in equation (3.2)

[2]p-value from de Long's test of 2 ROC curves for difference in discriminative ability of the score when applied to males and females separately

Considering the results from the modified PRSice, to allow inclusion of the ambiguous SNPs, the association p-value from the model containing all 15 SNPs was 2.72×10^{-160}. This too agrees with the R score using the dosage data, which is just a little more significant (**Table 3.7**). For the 15 SNPs the relative performance of each method is the same as was seen for 12 SNPs. The PRSice score gives a slightly lower performance for

males compared to R, and slightly higher performance for females, leading to a less significant p-value for the sex difference in PRSice.

Comparing the 2 PRSice scores allows the effect of removing the ambiguous SNPs to be evaluated. When the overall AUC of the score containing the 12 non-ambiguous SNPs, 0.782 (0.763, 0.801), was compared against the 15 SNP score (AUC 0.787 (0.768, 0.806)) it was found to be significantly lower with a p value of 0.0407 between the 2 ROC curves. In other words, inclusion of the 3 ambiguous SNPs increases the overall predictive accuracy of the GRS.

3.3 Discussion

3.3.1 GRS using R

From the above exploration, there is no evidence of a sex-based difference in the genetic risk profile for IPF based on 15 known association loci and meta-analysis effect estimates. Though *MUC5B* has a large impact on the AUC it does not appear to be driving the overall conclusions.

However, the number of female cases is low (196) making up only 24.7% of cases and the power to detect sex differences based on the UUS data sample size is likely to be low.

The discriminative ability of the overall GRS score in this case-control setting is moderate at 0.787 (0.768, 0.806). However, it is clear from the low positive predictive values that further improvements need to be made for any score based on genetics alone to be useable in a wider population setting.

There are many more susceptibility variants that contribute to the overall genetic architecture of IPF, albeit it with modest overall effect (section 2.5). If there were differences in the genetic risk profiles between sexes, then we simply haven't had the power in current GWAS for variants that are more associated with IPF risk in females to achieve genome-wide significance. In other words, we need to look beyond just the known genetic associations.

3.3.2 GRS using PRSice

The performance of the GRSs produced by manual coding and those produced by PRSice v1.25 have extremely good agreement when the scores contain identical SNPs, with any residual differences explainable by the difference between the use of dosage data during manual coding, and hard call genotypes for PRSice.

From the GRS sensitivity analysis, the effect of the ambiguous SNPs seems sizeable both in terms of model fit and discriminative ability of the score. However, the exclusion of the 3 ambiguous SNPs doesn't change the overall conclusions in terms of a sex difference, with no evidence of a sex stratified difference in either the score made up of 12 SNPs or, as seen earlier the score with 15 SNPs.

Further, the effect of ambiguous SNP exclusion may be negligible when creating the PRS based on genome-wide SNPs. In general, PRSs are created with many more SNPs than just the 15 used here. In addition, ambiguous SNPs are removed before clumping, so if within a clump the SNP with the highest association would have been an ambiguous SNP, the non-ambiguous SNP with the lowest p-value is chosen instead. This is likely to be in high LD with the removed SNP and may therefore act as a

reasonable proxy. Additionally, ambiguous SNPs are all transversion mutations – where a purine base is substituted by a pyrimidine or vice versa (see **Figure 1.1**). These occur less frequently than transitions – which are pyrimidine-pyrimidine or purine-purine substitutions. Transversions represent a bigger change within DNA at the molecular level and in exons are more likely to result in an amino acid change in the resulting protein[97] so important information may still be removed by excluding ambiguous SNPs.

This will be investigated further as a sensitivity for the PRS scoring when the scoring is expanded.

3.4 Summary

In this chapter it has been shown that a genetic risk score based on the 15 known genetic risk variants for IPF is significantly associated with IPF. There is no evidence of a sex difference in the discriminative ability or predictive accuracy of this score.

Though the AUC is moderate, the genetic risk score with 15 variants has low clinical utility and is not suitable for screening at population level.

Lastly a technical validation of PRSice was performed and found to agree with the manually produced GRS. With this, a benchmark had been created to use when evaluating further scores that will also be calculated using PRSice. However, there is a potential shortcoming in PRSice with the removal of ambiguous SNPs that will be investigated further.

Next we consider extending the analysis to consider scores with many more variants at multiple p-value thresholds.

4 PRS WITH STANDARD GWAS

With no sex difference when the known risk variants for IPF are combined into 1 score, the scoring is expanded to capture more of the phenotypic variation. In this chapter PRS are calculated for a range of p-values using the full set of available sex-agnostic GWAS results.

The scores are then assessed for sex-specific performance. When doing this the initial focus is on the model fit and identifying where the addition of a sex interaction term improves the fit. The discriminative ability is investigated in detail and the p-value threshold where the sex difference in discriminative ability is largest as well as the p-value thresholds at which the AUC is maximised for each sex individually are identified.

4.1 Methods

Analyses were performed using R version 4.1.0[92], PLINK 2.0[98], PRSice v 1.25[80] with programs developed and executed on SPECTRE, the University of Leicester's HPC cluster.

4.1.1 Data

Base data

The full set of GWAS summary statistics results comprises 10,790,934 variants in text format. Variants were included if they had a minor allele count of at least 10. For each variant, the information described in section 3.1.1 was present. Additionally, the file contained the effect allele frequency (across the included studies) and study specific

information including an indication of which of the 3 studies contributed to each variant and the study specific imputation quality.

Target data

The target data were the UUS data as described in section 3.1.1 in PLINK format. For this analysis, only genotyped and well imputed genotypes (see section 3.1.1 for details) with an $R^2 > 0.5$ were selected, giving a total of 26,309,641 possible variants, with variants not excluded based on minor allele count.

4.1.2 Data preparation

First sex and phenotype were added to the PLINK files using the information available in 3.1.1. A covariate file was created containing sex and the first 10 principal components for each individual. Variables in the base data were renamed to match with PRSice expectations of the column names and the rsid column was re-coded to contain chromosome:position to allow matching of the SNPs between the GWAS and the PLINK data files.

Next, the data were prepared so that only SNPs common to both base and target data were considered for analysis with redundant variants removed from the target data. 10,787,395 common variants were retained. This reduces the size of the files and speeds up polygenic scoring, particularly the clumping.

4.1.3 Thresholding of PRS scores

PRSice was set up to include sex and the 10PCs as covariates, fitting a model (3.2) to assess the association between score and phenotype. Default clumping options were

used. Scores for all thresholds were output to allow further analysis and exploration. The liability R^2 was requested using the highest available estimate of IPF prevalence in the general population of 63 cases per 100 000. For a representative PRSice call with all options used, see APPENDIX C.

An iterative process was used to identify the range of thresholds for PRS scoring. First, PRS scores were created for the default thresholds of 0.001, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5. As PRS scoring is computationally intensive, further low-resolution scoring was performed to cover the range towards a p value of 1 with individual scores calculated for pT 0.6 to 1 in increments of 0.1. Several more thresholds were also added pragmatically to cover thresholds < 0.001 and in areas where big changes in overall score association were seen in the original default bar chart. Finally, high-resolution scoring was performed on the range of interest.

4.1.4 Assessment of performance

The following was assessed at each high-resolution threshold:

- Association of score with phenotype using model *(3.2)* (performed by PRSice). Model fit was evaluated by a LR test of the 'null model' which only includes covariates in the linear predictor compared to the model including score and all covariates.
- Full model fit when sex interaction is included *(3.3)*. Evaluated by LR test of model only including covariates in the linear predictor compared to the full model including score, sex interaction and all covariates.

- P-value of the sex interaction – as determined by LR test of models with and without the sex interaction term, *(3.3)* compared to *(3.2)*.
- AUC and confidence interval stratified by sex.
- Test for difference in sex-specific AUCs using de Long's test.

The above was visualised by plotting across the range of thresholds used in the high-resolution scoring, with the aim of characterising the overall difference in distribution between the sexes. As for the PRS scoring, to correct for multiplicity, the significance threshold was set at $p < 0.001$.

The score with largest sex difference in discriminative ability (AUC) was identified. For this score the following was assessed:

- ROC plot and kernel densities by sex
- Liability R^2 for this threshold, based on prevalence 1.25 to 63 per 100 000
- Quantile plot by sex

The model with the most significant sex-interaction p-value (model *(3.3)*) was also identified and described.

4.1.5 Optimising PRS score by sex

To evaluate if sex-specific PRS scores have additional clinical utility, the best model with the highest overall predictive accuracy was determined overall and by sex. For this the target data were split by sex and the PRS optimised for each sex independently. As there are fewer SNPs at extremely low p-values, lower resolution scoring was possible for the p-value range from $5x10^{-8}$ and $5x10^{-4}$. These optimised

scores were considered in turn and the sex difference tested for difference within and between these optimum scores (**Figure 4.1**).

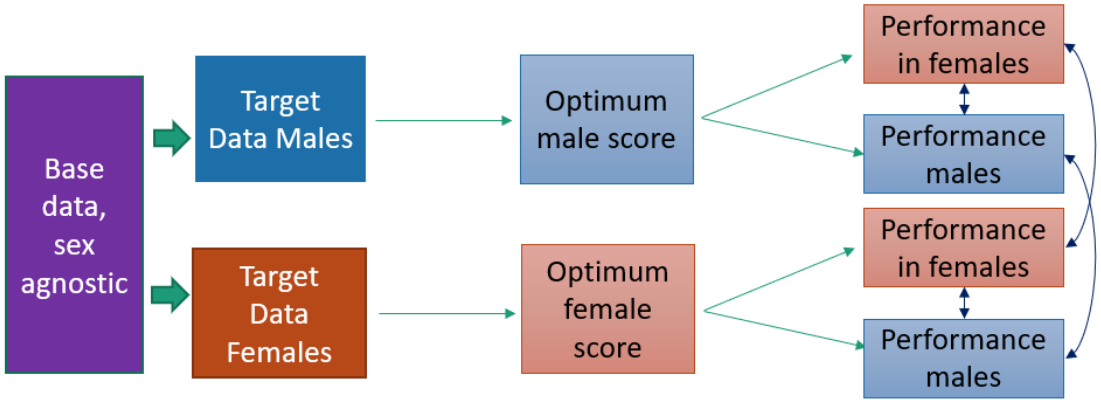


Figure 4.1: Creation and testing of optimum scores by sex

4.1.6 Sensitivity - Ambiguous SNP inclusion

Using the 3 SNPs excluded during the GRS replication, a check was performed on the retained SNPs within those clumps in the full analysis. The LD proxy tool[99] from LDLink[100] was used to examine the correlation between the removed ambiguous SNP and the top variant used instead to identify if the non-ambiguous SNP chosen as the top variant was an appropriate proxy for the removed ambiguous SNP.

As a sensitivity to the main analysis, the high-resolution scoring was repeated using the modified PRSice to allow inclusion of the ambiguous SNPs.

4.1.7 Sensitivity – Minor Allele Frequency

Due to the low sample size, especially within the female cases, it is possible that sex differences seen are due to false positives (type I error). The analysis was repeated with only SNPs with a MAF > 5% in the target data. A MAF of 5% means the variant

allele is present in at least 1097 alleles for a SNP if all individuals are genotyped (measured across the study sample in general).

4.1.8 Sensitivity – Model checking for outliers and influence

For the score with the biggest sex difference, the deviance residuals, leverage, and influence from the logistic regression were used to identify if there were any individual subjects that drive the results of the PRS score association.

4.2 Results

A total of 9,169,326 non-ambiguous variants were included. After clumping, 1,101,670 predictors were loaded for polygenic scoring.

4.2.1 Thresholding of PRS scores

The PRS which produces the strongest association between score and phenotype is at P_T 0.001 (**Figure 4.2**). This is the most stringent threshold tested here, but is much less stringent than the genome-wide significance level of $5x10^{-8}$. Note that the p-value of this 'best score' is $4.2x10^{-53}$, considerably lower than the 9.11×10^{-156} obtained from the score containing the 12 non-ambiguous known risk variants (**Table 3.7**).

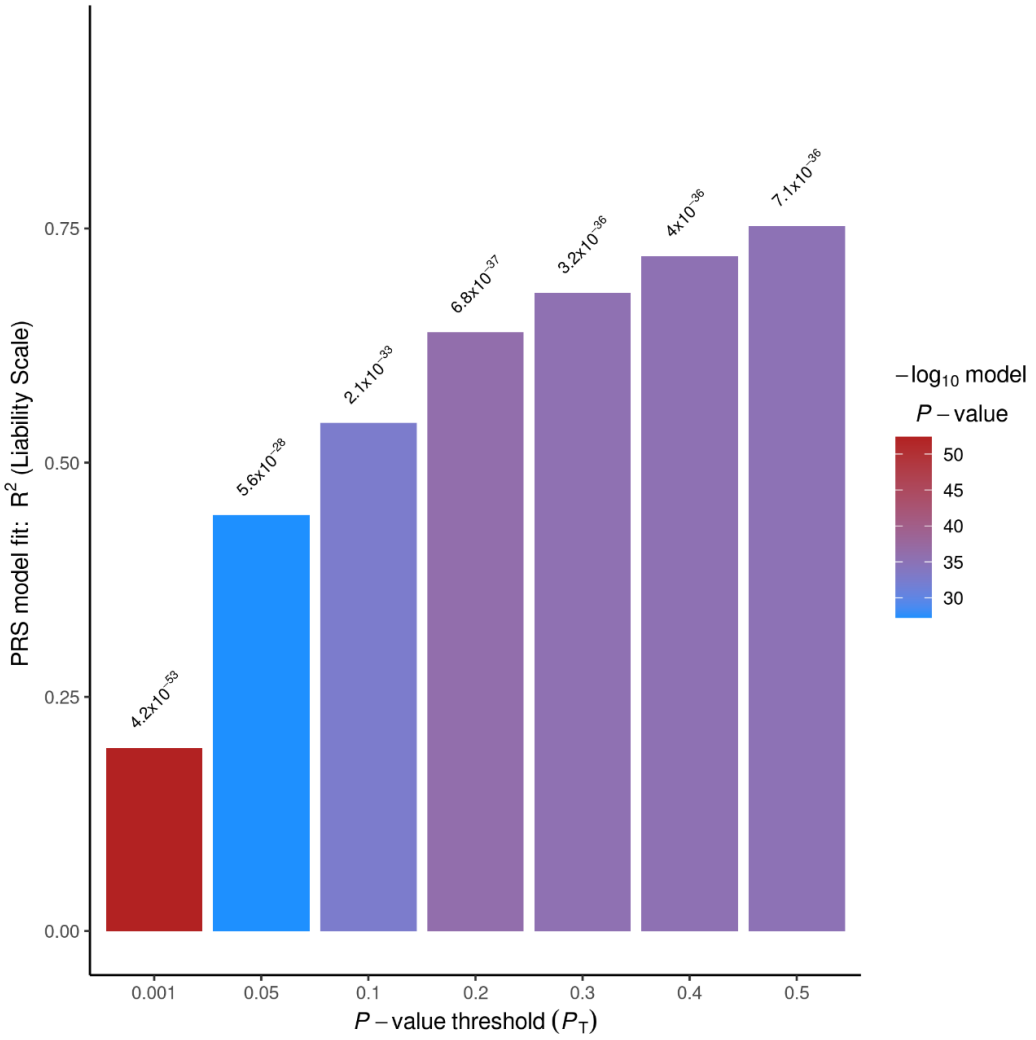


Figure 4.2: Model fit of low-resolution default thresholds.

R^2 on the liability scale assuming a population prevalence of 63 cases per 100,000

The analysis of association, the sex interaction, and ROC curve AUC results are summarised in **Table 4.1** for these initial thresholds. As might be expected from the overall p-value of the model, the AUCs achieved are lower compared to the analysis of the known risk variants. The AUCs of the female ROC curve was higher 0.685 (95%CI 0.648, 0.722), compared to the male curve 0.642 (0.619, 0.664), with this difference nominally significant at the 5% level (p=0.047), but not at the adjusted 0.1% level.

	ROC AUC			association	
Threshold	male (95% CI)	female (95%CI)	Pval[1]	Score pval[2]	sex*score Pval[3]
0.001	0.642 (0.619, 0.664)	0.685 (0.648, 0.722)	0.047	$4.2x10^{-53}$	0.195
0.05	0.584 (0.560, 0.608)	0.638 (0.601, 0.674)	0.017	$5.6x10^{-28}$	0.019
0.1	0.593 (0.570, 0.616)	0.658 (0.623, 0.694)	0.003	$2.1x10^{-33}$	0.008
0.2	0.604 (0.581, 0.627)	0.654 (0.618, 0.689)	0.022	$6.8x10^{-37}$	0.036
0.3	0.606 (0.583, 0.629)	0.651 (0.615, 0.687)	0.039	$3.2x10^{-36}$	0.065
0.4	0.607 (0.584, 0.629)	0.648 (0.612, 0.685)	0.057	$4.0x10^{-36}$	0.089
0.5	0.607 (0.584, 0.629)	0.648 (0.612, 0.684)	0.057	$7.1x10^{-36}$	0.092

Table 4.1: Summary of Results from initial default PRS thresholds.

[1]p-value from de Long's test of 2 ROC curves

[2]Score p-value as generated by PRSice, from logistic regression including score + covariates

[3]p-value from likelihood ratio test of model including sex interaction against base model

Further scores were calculated at the following thresholds: pT 0.6 to 1 in increments of 0.1; additional thresholds < 0.001 ($5x10^{-8}$, $5x10^{-7}$, $1x10^{-6}$, $5x10^{-6}$, $1x10^{-5}$, $5x10^{-5}$, $1x10^{-4}$, $5x10^{-4}$); and between pT0.001 and pT0.1 (0.005; 0.01 - 0.09 in increments of 0.01). Using all scores calculated up to this point, the AUCs and 95% confidence intervals were plotted by sex to visualise the relationship between predictive accuracy (ROC curve AUC) and sex as the threshold for PRS score inclusion increases (**Figure 4.3**). The log scale illustrates the incredibly close performance at more stringent p-value thresholds, consistent with results from the GRS (chapter 3). As the threshold increases there is a divergence in performance between the sexes, which levels off at around pT > 0.3.

High-resolution scoring was performed with p-value thresholds from 0.0005 to 0.25, incremented by 0.0001.

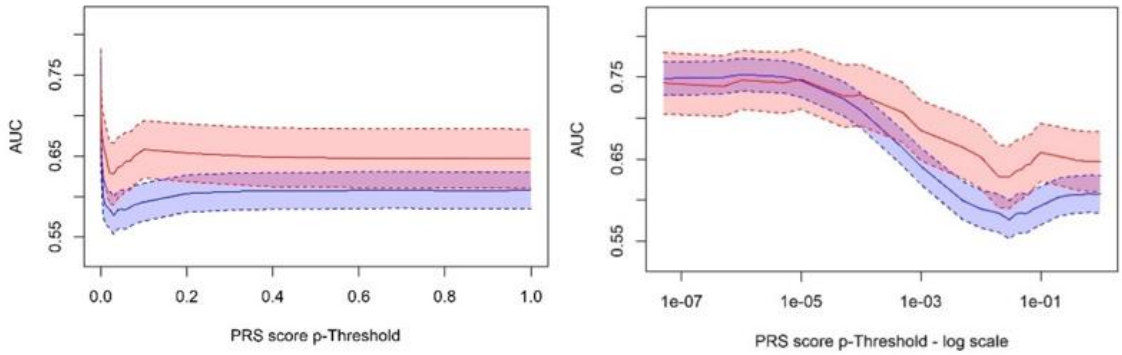


Figure 4.3: AUC by Sex across p-value thresholds, low-resolution scoring. Linear x-axis (right); Log x-axis (left). Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI.

4.2.2 Assessment of overall performance – PRSice model fit

Model fit with and without sex interaction is shown in **Figure 4.4**. As the p-value threshold (pT) for including SNPs in the PRS increases there is a sharp initial drop in the model p-value. The lowest fit occurs around a pT of 0.0296. The association at the threshold with the poorest fit is still considered high at 3.33×10^{-24}. As the pT increases beyond this point, the strength of the score association increases gradually to around $P = 1 \times 10^{-36}$. No real improvements are seen in the association of the overall score with phenotype after pT 0.15.

4.2.3 Optimisation of sex differences

Sex interaction model fit

The model fit with and without sex interaction follows the same pattern. There are areas of the graph where the sex-interaction line is visible above the default line (between pT 0.05 and 0.10) and areas where it is below (pT 0.2 onwards). **Figure 4.5**

shows the p-value of the sex interaction term when it is added to the model for each pT.

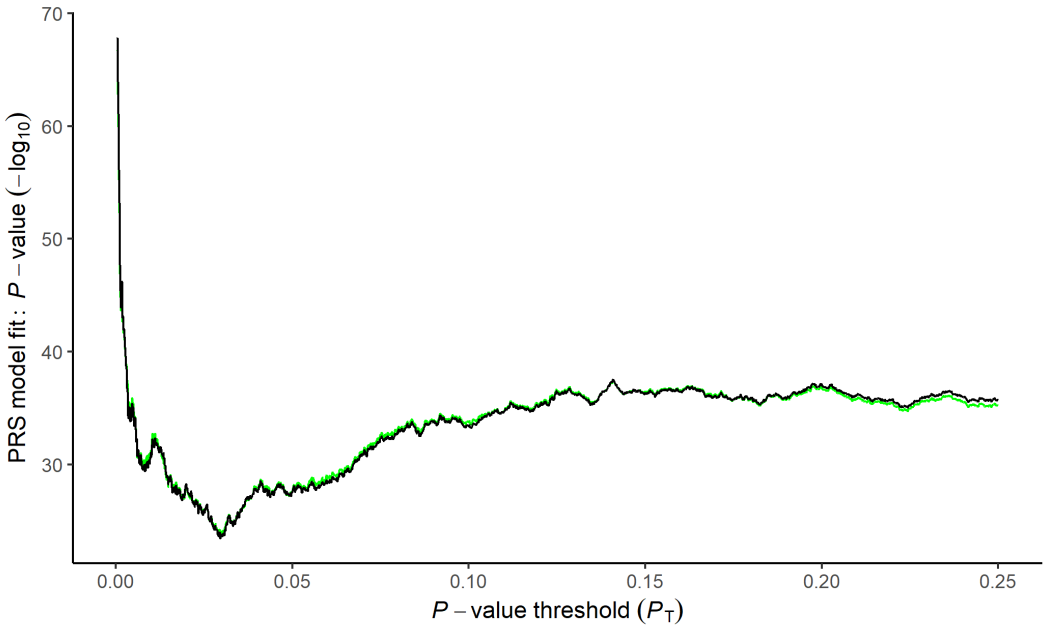


Figure 4.4: PRS model fit across p-value thresholds.
Black line shows model fit with score and covariates; green line shows the same model with additional sex interaction. P-values of model fit from likelihood ratio tests against model with just covariates.

At the lowest thresholds, the addition of the sex interaction term does not improve the model and the difference in score between cases and controls is the same in males and females (also seen with the GRS in chapter 3). By a pT of 0.003 the sex interaction has increased in significance to 0.003, with the smallest p-value at a pT of 0.0033 (p=0.0034) (**Table 4.2**). This does not reach corrected significance (red line, **Figure 4.5**). The sex interaction association then weakens (higher p-value) to p=0.07 at a pT of 0.0135. As scores within PRSice are not standardised for number of SNPs in the score, the coefficients of the sex interactions cannot be readily compared between thresholds.

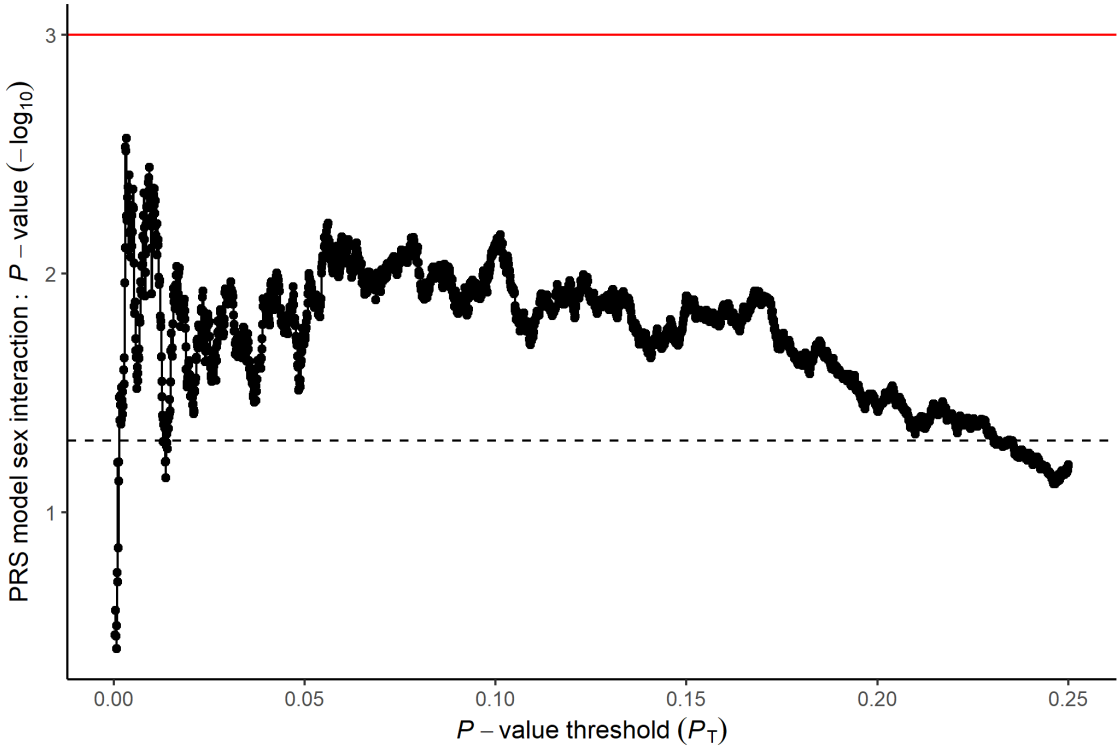


Figure 4.5: P-value of sex interaction term across p-value thresholds. Sex interaction p-value from LR test. Black reference line at p=0.05, red line at p=0.001. Y axis presented on -10log scale.

Sex difference in discriminative ability

The discriminative ability of the score across thresholds, by means of the AUC and 95% confidence intervals by sex, is shown in **Figure 4.6** and the corresponding p-values from the test of the difference between the 2 curves in **Figure 4.7**. The shape matches that seen during the low-resolution scoring (**Figure 4.3**), as expected and the results mirror the sex-interaction results. There is a small range of low pTs where a divergence is seen between male and female predictive accuracy. The discriminative ability then re-converges to some extent – seen by overlap in the confidence intervals in **Figure 4.6**. In the range of pTs from around 0.05 to 0.2 very little overlap is seen between the sex-specific predictive accuracies. For the range of thresholds shown, the female

predictive accuracy is higher than the male predictive accuracy. The wider confidence intervals for the female estimates are reflective of the relative sample sizes.

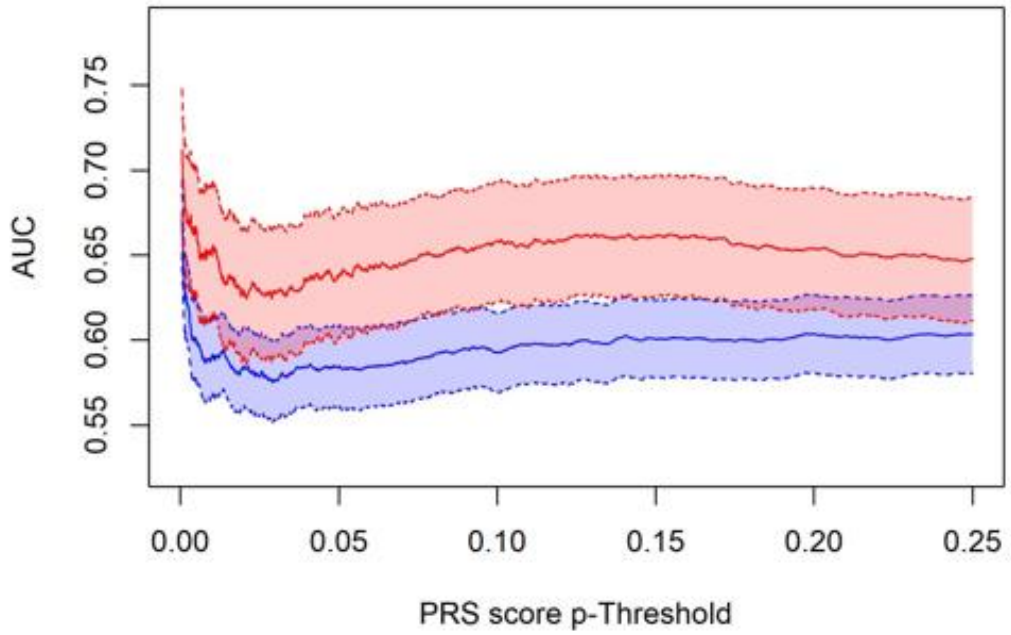


Figure 4.6: AUC by Sex across p-value thresholds, linear scale, high-resolution. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI

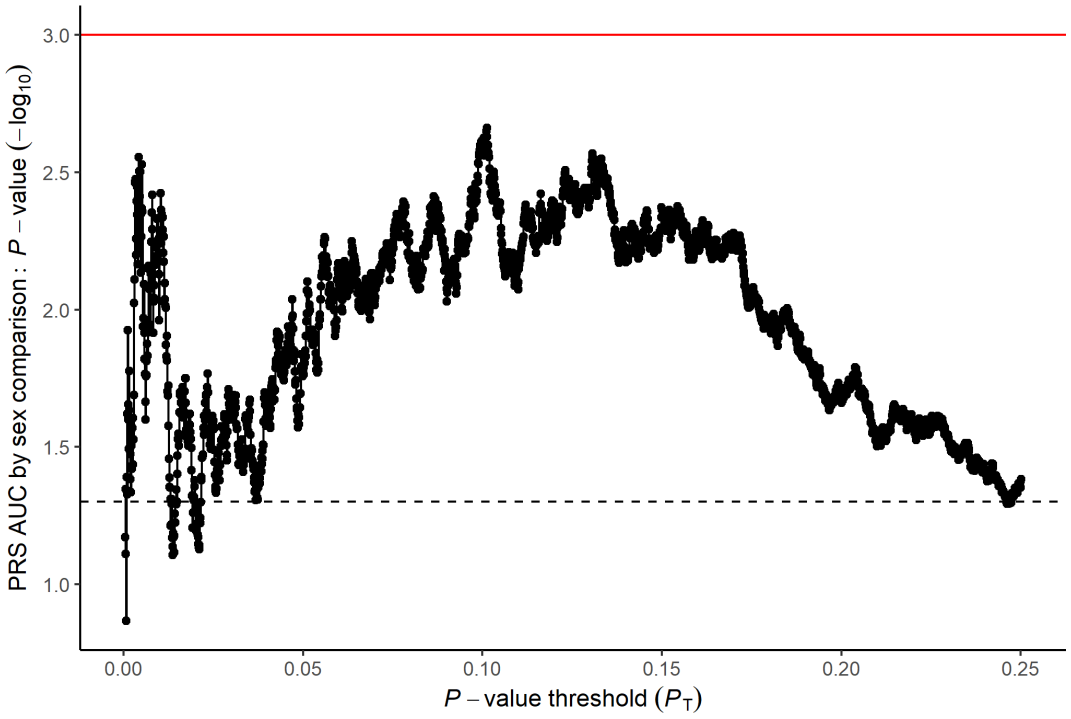


Figure 4.7: Comparison of AUC by Sex across p-value thresholds, Y axis shows p-value from de Long's test for uncorrelated ROC curves. Black reference line at p=0.05, red line at p=0.001. Y axis presented on -10log scale.

Threshold with largest sex difference

The threshold with largest difference in AUC occurs at pT 0.1012, doesn't reach the corrected significance threshold of 0.001. The five top scores are presented in **Table 4.2**.

		ROC AUC			association	
Threshold	N SNPs	male (95% CI)	female (95%CI)	Pval[1]	Score pval[2]	sex*score Pval[3]
Biggest AUC diff						
0.1012	203764	0.593 (0.570, 0.616)	0.659 (0.624, 0.695)	0.0022	$2.43x10^{-34}$	0.0069
0.1011	203588	0.593 (0.570, 0.616)	0.659 (0.624, 0.695)	0.0022	$2.61x10^{-34}$	0.0070
0.1013	203927	0.593 (0.570, 0.616)	0.659 (0.623, 0.694)	0.0024	$2.27x10^{-34}$	0.0072
0.1004	202336	0.593 (0.569, 0.616)	0.658 (0.623, 0.694)	0.0024	$4.91x10^{-34}$	0.0071
0.1005	202513	0.592 (0.569, 0.616)	0.658 (0.623, 0.694)	0.0024	$5.63x10^{-34}$	0.0076
Strongest sex-interaction						
0.0033	9443	0.606 (0.583, 0.629)	0.672 (0.635, 0.710)	0.0034	$5.52x10^{-22}$	0.0027

Table 4.2: Thresholds with largest difference in performance between the sexes

[1]p-value from de Long's test of 2 ROC curves

[2]Score p-value as generated by PRSice, from logistic regression including score + covariates

[3]p-value from likelihood ratio test of model including sex interaction against base model

At pT 0.1012 203,764 SNPs are included in the score. This comprises 18.5% of the SNPs considered for the PRS scoring. The liability R^2 at this threshold is estimated to be in the range 33.96 – 54.57%.

The ROC curve at pT 0.1012 (**Figure 4.8**) shows the sex difference, with the score performing better for females compared to males across the full range of sensitivity and specificity, despite not meeting the more stringent significance threshold of 0.001.

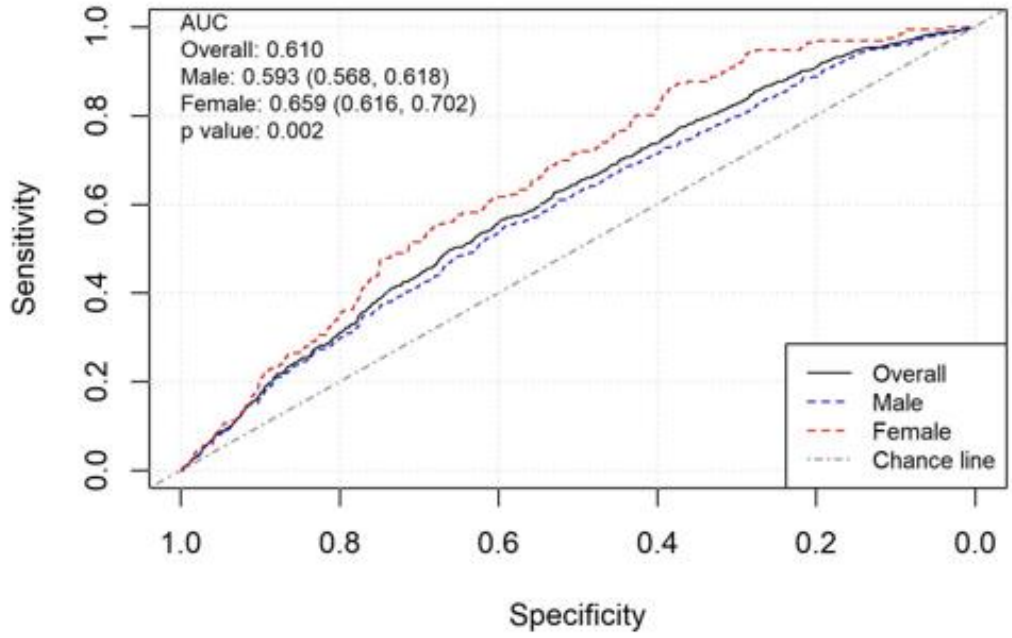


Figure 4.8: ROC curves by sex of PRS at p-value threshold 0.1012. P-value from de Long's test for uncorrelated ROC curves

When examining the distribution of cases and controls by sex at pT 0.1012 (**Figure 4.9**) the difference between cases and controls is less pronounced compared to the GRS (chapter 3), in line with the lower AUCs observed here. The distribution of male cases lies to the left and closer to controls than the female distribution.

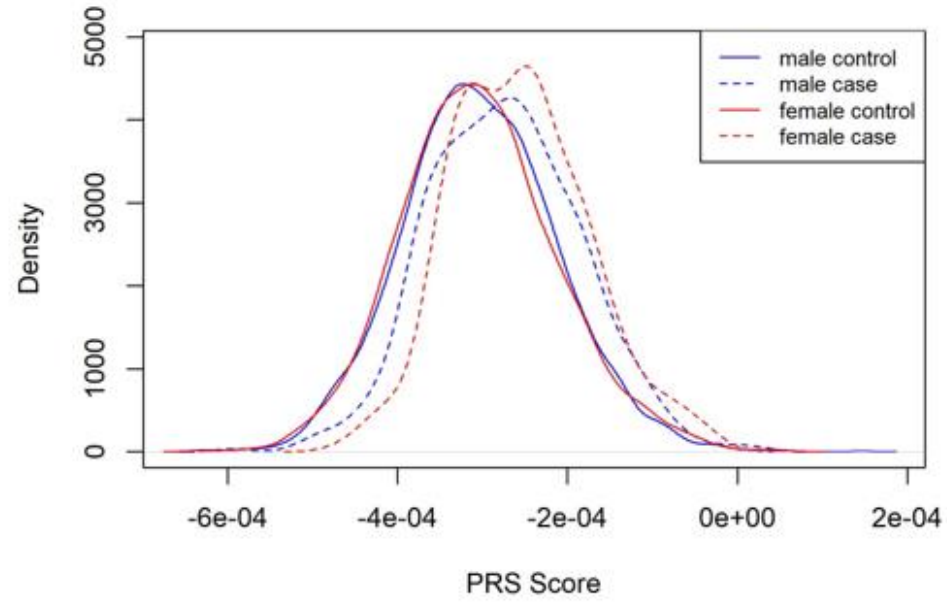


Figure 4.9: PRS distribution by disease status and sex at p-value threshold 0.1012

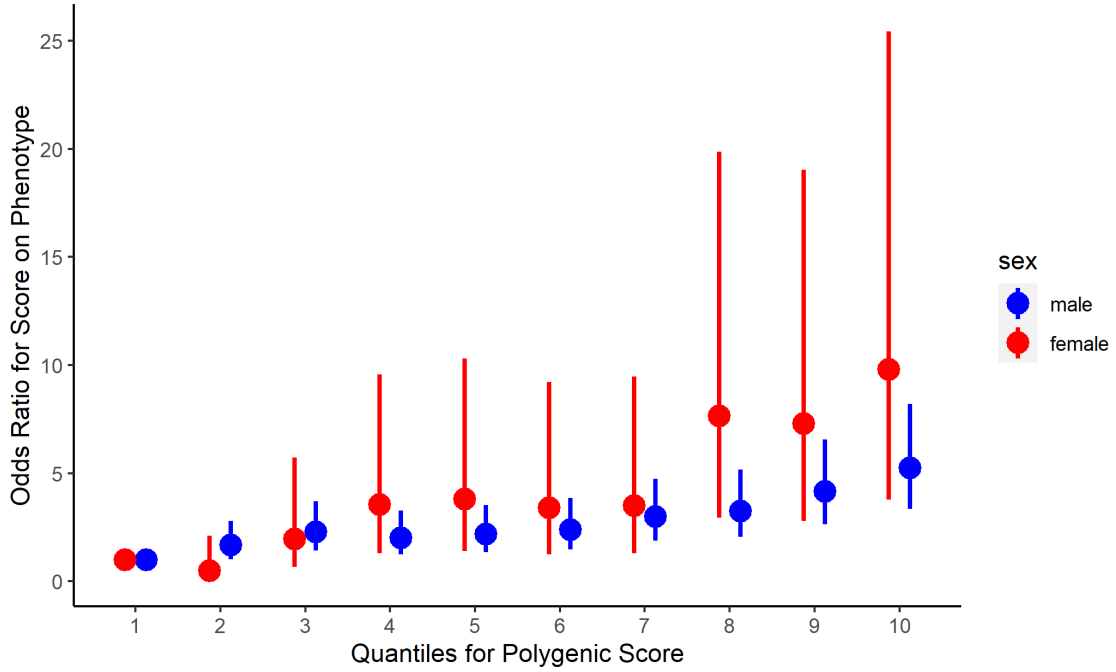


Figure 4.10: Quantiles by sex, for PRS at p-value threshold 0.1012

Figure 4.10 shows the performance of the score by splitting the score into deciles by sex and considering the relative odds ratio of IPF compared to the first decile. Increasing score corresponds to increasing odds of IPF with an effect that is more pronounced in females, though at each decile the confidence intervals overlap.

4.2.4 PRS optimised by sex

From **Figure 4.3**, the performance was best at more stringent p-value thresholds, with very little discernible difference between the AUCs. The score most associated with IPF status overall occurs at a pT of $1x10^{-6}$, with 62 SNPs included in the score (**Table 4.3**). This is also the score most associated with IPF status for males. The score with the best discriminative ability in females has a slightly less stringent pT of $1x10^{-5}$ and contains 134 SNPs. When the best male pT is applied to each sex in turn, the AUC is higher in males than females, but not significantly so (**Table 4.3**).

Whilst male performance reduces, though not significantly (p=0.044), between the thresholds of $1x10^{-6}$ and $1x10^{-5}$, the female performance shows no difference between these two thresholds (p=0.907) (**Table 4.3**).

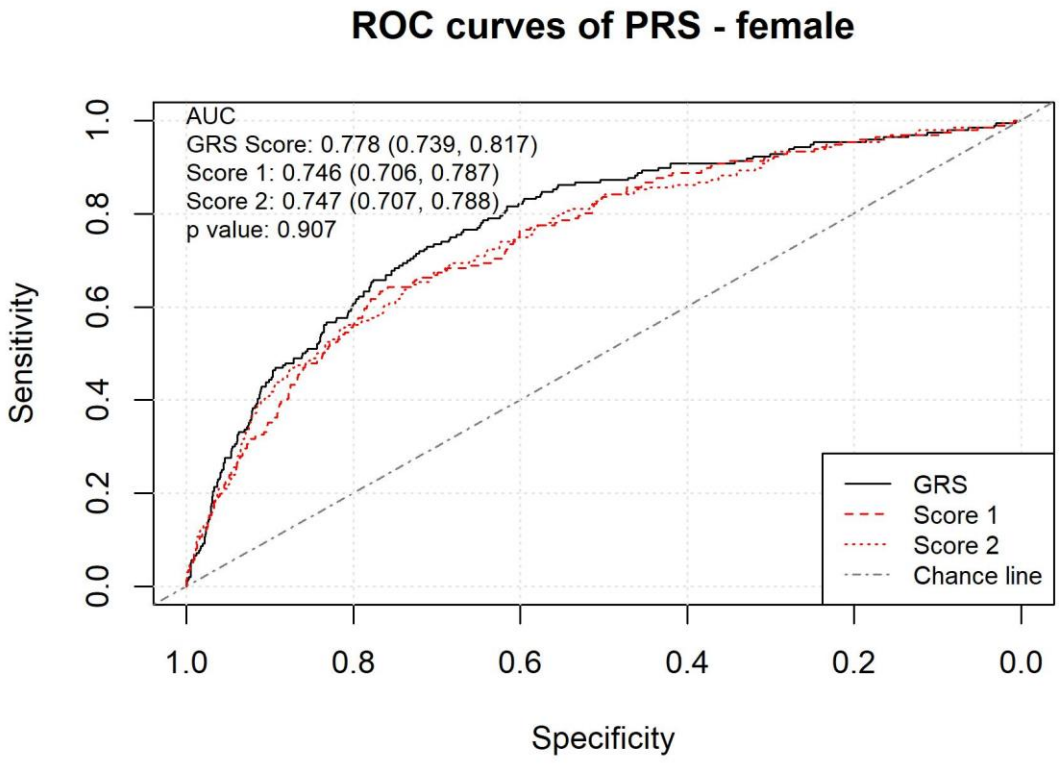
		ROC AUC			association pval[2]	
Threshold		male (95% CI)	female (95%CI)	Pval[1]	male	female
best male	$1x10^{-6}$	**0.753** (0.730, 0.776)	0.746 (0.706, 0.787)	0.749	$1.2x10^{-93}$	$1.5x10^{-32}$
best female	$1x10^{-5}$	0.744 (0.721, 0.768)	**0.747** (0.707, 0.788)	0.931	$3.3x10^{-89}$	$4.4x10^{-34}$

Table 4.3: Relative Performance of best Male and Female thresholds

[1]p-value from de Long's test of 2 ROC uncorrelated curves

[2]Score p-value as generated by PRSice, from logistic regression including score + covariates. Produced by PRS scoring by each sex separately

The GRS (12 non-ambiguous SNPs with known association) from chapter 3 also outperforms both sex-specific thresholds as seen by the black ROC curves in **Figure 4.11**.



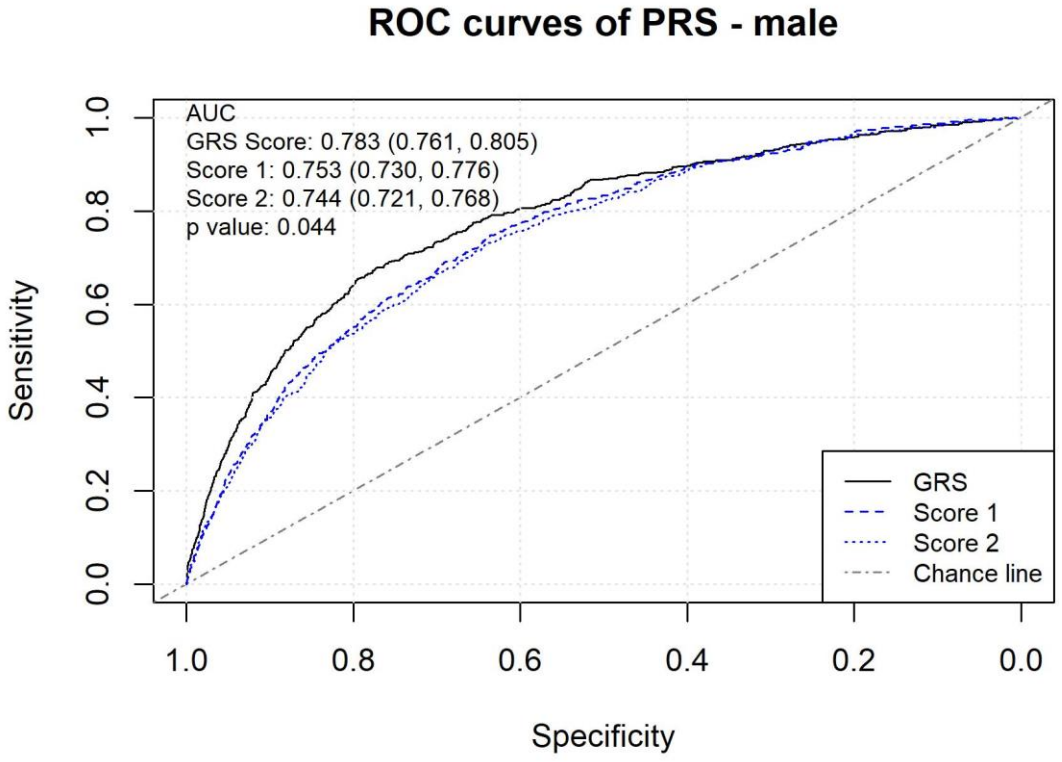


Figure 4.11: ROC curves of best thresholds by sex. Score 1= best Male score at pT 1x10-6, Score 2= best Female score at pT 1x10-5, p-values in each panel de Long's test of correlated ROC curves between score 1 and 2.

Though there is little difference in the AUC by sex for the best overall scores at pT $1x10^{-6}$, the quantile plot by sex (**Figure 4.12**) does show some differences. For males, the risk of IPF for the subjects with the highest 10% of PRS scores is 35.67 (CI: 17.44, 72.94) times that of those in the lowest 10% of scores. For females, though the confidence interval overlaps with the male estimate, this OR is much smaller at 15.67 (CI: 6.67, 36.84).

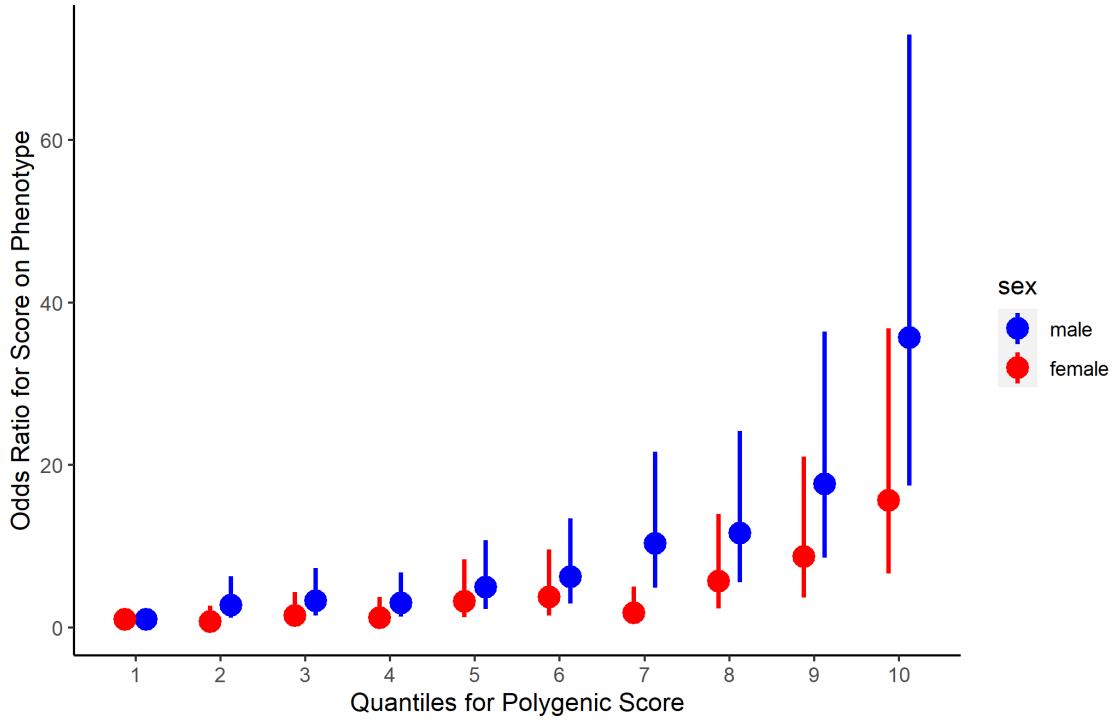


Figure 4.12: Quantile plot by sex for PRS score at pT $1x10^{-6}$

4.2.5 Sensitivity – Ambiguous SNPs

The 3 Ambiguous GRS SNPs and the proxy used for each when these ambiguous SNPs were removed by PRSice are shown in **Table 4.4**. Matches for 3:169481271 and

3:169481271 are exceptionally high, both in terms of LD, but also the beta and p-value. The match with rs78238620 (3:44902386) is slightly less good, with a beta coefficient that differs at the 2nd decimal place and an R^2 of 0.8525.

Ambiguous SNP			Proxy Chosen			LD
SNP	beta	p-value	SNP	beta	p-value	R^2
3:44902386	0.4594	$5.12x10^{-10}$	3:44845649	0.4755	$7.74x10^{-10}$	0.8525
3:169481271	0.2668	$7.09x10^{-13}$	3:169486144	0.2655	$9.75x10^{-13}$	0.9949
13:113534984	-0.2643	$1.34x10^{-10}$	13:113540425	-0.2636	$1.58x10^{-10}$	0.9823

Table 4.4: Proxies for Selected Ambiguous SNPs. LD based on all European reference populations available in LDLink

When the modified PRSice was used to include ambiguous SNPs, more variants were included in the PRS scoring after clumping. 1,166,017 partially independent SNPs were loaded, compared to 1,101,670 in the primary analysis, an increase of 64,347.

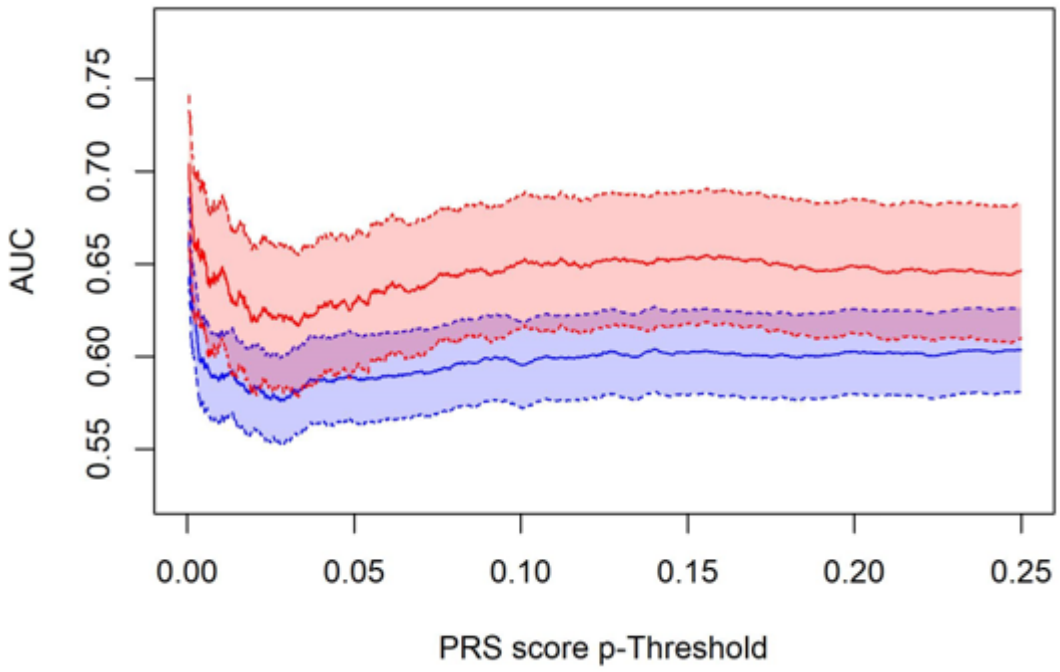


Figure 4.13: AUC by Sex across p-value thresholds. Sensitivity - ambiguous SNPs included.

There is more overlap between the male and female AUCs and the AUCs of both sexes are a little lower, but the overall picture is consistent with the primary analysis (**Figure 4.13**) and the female discriminative ability remains higher across the range of pT. In line with the overlapping AUC CIs, the p-value of the difference in AUCs s less significant and the sex interaction term had a less pronounced effect when added to the logistic regression model (*3.3*).

The AUC with the maximum sex difference occurs at pT 0.0041 and comprises 12,493 SNPs. This threshold, though not in the top 5, also showed a large sex difference in discriminative ability in the primary analysis (where it included 11,352 SNPs).

4.2.6 Sensitivity - Minor Allele Frequency

When only variants with a MAF > 5% were kept, 4,543,204 variants remained for consideration, reducing to 150,500 predictors after clumping. Comparing this to the 1,101,670 in the primary analysis, highlights the large proportion of low frequency variants present in the data.

Figure 4.14 shows higher AUCs for both sexes at all thresholds, with the female AUC now stabilising at around 0.70, and more separation between the sexes compared to the primary analysis. The threshold with the biggest sex difference at pT 0.1193 shows a significant difference (p=0.0004) and contains 42,165 SNPs (**Table 4.5**). Compared to the primary analysis, where the score with the largest difference contained 203,764 SNPs at 0.1012, the liability R^2 is lower, with up to 17.15% of variance explained (estimate for prevalence of 63 per 100 000).

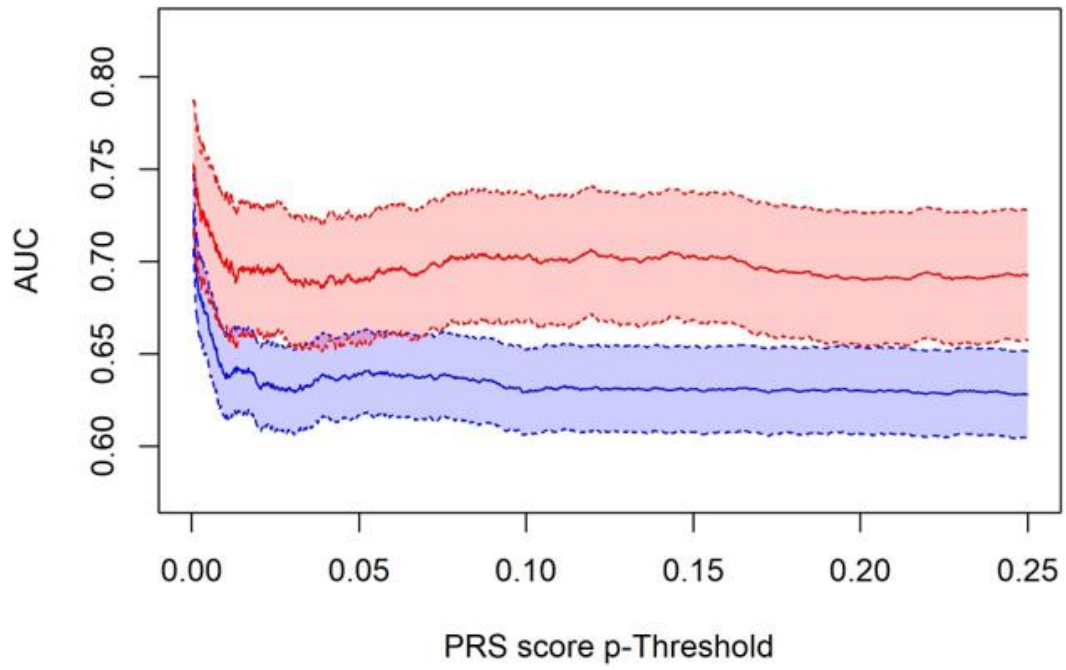


Figure 4.14: Sensitivity - MAF > 5% in target data

4.2.7 Outliers and Influence

Investigation of the deviance residuals for the logistic regression model using a PRS score with p-value threshold of 0.1012 (APPENDIX , **Figure D.1**) did not identify any extreme outliers for investigation.

The leverage (h_{ii}) and Cook's distance both indicated that one subject was influential in the fit of the logistic regression model that used the PRS score created from a p-value threshold of 0.1012 (APPENDIX , **Figure D.2**). This influential record was a female IPF case with a score of -3.15×10^{-4}, well below the median score for female cases (see **Figure 4.9**). When the performance of the score was re-assessed following exclusion of this subject, the female performance increased slightly though not significantly (compared to the primary analysis) with a corresponding relative sex difference increase at this threshold (**Table 4.5**).

		ROC AUC			association	
					Score	sex*score
Sens	Thresh	male (95% CI)	female (95%CI)	Pval[1]	pval[2]	Pval[3]
Primary	0.1012	0.593 (0.570, 0.616)	0.659 (0.624, 0.695)	0.00218	$2.43x10^{-34}$	0.0069
Amb	0.0041	0.597 (0.574, 0.620)	0.658 (0.621, 0.696)	0.00646	$1.01x10^{-33}$	0.0096
MAF > 5	0.1193	0.631 (0.608, 0.654)	0.707 (0.672, 0.741)	0.00039	$4.35x10^{-46}$	0.0021
Excl outlier	0.1012	0.593 (0.568, 0.618)	0.660 (0.617, 0.703)	0.00192	$2.36x10^{-34}$	0.0059

Table 4.5: Sensitivity summary - thresholds with largest difference in performance between the sexes

[1]p-value from de Long's test of 2 ROC uncorrelated curves

[2]Score p-value as generated by PRSice, from logistic regression including score + covariates

[3]p-value from likelihood ratio test of model including sex interaction against base model

4.3 Discussion

Extremely strongly associated PRS scores were produced at low p-value thresholds with the overall best score overall in this chapter coming from just 62 SNPs at a p-value threshold $1x10^{-6}$. However, both the association and discriminative ability were lower than those seen in Chapter 3 with the GRS of known IPF risk variants. This will be considered in the overall discussion (Chapter 6). Unlike the strength of association of the PRS, the liability R^2 continues to increase as pT increases, as a larger proportion of variation is captured (shown in **Figure 4.2**). This is as expected, and it is possible for score that captures all genetic variation to perform poorly at predicting phenotype[77].

4.3.1 Sex differences

The first aim was to investigate the existence of any sex differences in the genetic architecture of IPF and characterise these.

In the event of a genetic risk difference between sexes, variants that predominantly affect females will not have reached overall genome-wide significance in the GWAS performed to date. These effects can be captures in a PRS as the p value threshold for

is increased from genome-wide significance towards 1. As SNPs were added and the p-value threshold became less stringent, the relative predictive ability in females became better compared to males, with the largest difference at pT 0.1012. The difference in AUC between the sexes had a p-value of 0.0022. Though it does not reach the corrected significance threshold of 0.001, the overall distribution of the AUCs with a divergence between the sexes at thresholds from 0.05 to 0.25, as well as the robustness of this distribution to sensitivity analyses indicates that further investigation is warranted. The sensitivity analysis including only MAF > 5% established that the observed sex difference was not being driven by variants rare in the target data. In fact, removing these increased both the predictive accuracy and sex difference, with the sex difference reaching statistical significance (p=0.00039).

There are a few possible explanations for the higher female predictive accuracy compared to males as the p-value threshold increased. It's possible that female specific risk variants are underpowered in GWAS performed to date and inclusion of them in a PRS with less stringent thresholds allows them to be included. Another relates to a possible sex-specific liability threshold, where, compared to male IPF patients, females need a higher number of risk alleles to develop the disease. A further possibility is the gender diagnostic bias, as females are less likely to be diagnosed with IPF, the average diagnosed female may have 'worse' disease than the average diagnosed male.

4.3.2 Sex-specific risk prediction

The second objective considers the clinical utility and whether incorporating sex can improve IPF risk prediction.

When considering clinical utility, the focus is on scores that have the best predictive accuracy and discriminative ability between cases and controls. Within the analyses performed in this chapter, considering sex-specific thresholds calculated from sex-agnostic weightings did not improve the sex-specific predictive accuracy.

4.4 Summary

In this chapter the base data from the IPF meta-GWAS was used to extend the GRS to include many more SNPs.

When performance of these scores was assessed by sex, divergence occurred as the p-value threshold approached 0.1 with female scores producing better discriminative ability compared to males. The difference did not reach significance when correction for multiple testing was considered. This general trend was robust to several sensitivity analyses and it was shown that considering just common variants improved both predictive accuracy and the observed sex difference with the sex difference significant at the 0.1% level.

The scores with the best discriminative ability between cases and controls did not show any sex differences when using the sex-agnostic weightings. In fact, the GRS (Chapter 3) outperformed any of the PRS scores calculated in this chapter.

Whilst we do have some evidence for sex-differences at less stringent pTs, there is no evidence to support sex-specific PRS scores to aid risk prediction. However, using sex-specific effect size estimates may be more informative. These will be considered next.

5 PRS WITH SEX INTERACTION GWAS

In this chapter a GWAS with sex interaction is considered as a means of creating sex-specific effect size estimates. This GWAS with sex interaction is conducted in the same dataset as the GWAS that was used as base data for the previous chapters is used.

The sex-interaction terms are meta-analysed and PRS scoring is applied to the sex interaction terms themselves to investigate if the cumulative effect of these sex interactions favours one sex. Then, to link it back to the analysis in the previous chapter and our second overall objective, we consider if using the sex-interaction term effect direction allows us to improve the predictive accuracy of a PRS score. Variants are partitioned into 'male specific' and 'female specific' and these variant sets are scored based on the main effects base data from the analysis in Chapter 4.

After that, the female effect estimates are investigated. These are meta-analysed, and a GWAS is performed to identify female specific associations that were not present in the sex agnostic GWAS.

Lastly, the predictive accuracy of PRS scores produced using the female estimates is explored and we consider if adding female specific hits to the GRS improves predictive accuracy.

5.1 Methods

5.1.1 Data

The original base data comprised 3 studies (UK, Chicago, and Colorado). These data were analysed using a more complex model that included additional terms for sex and a variant (SNP) by sex interaction as below in equation (*5.1*)

$$IPF\ Status_i = \beta_0 + \beta_1 SNP_i + \beta_2 Sex_i + \beta_3 SNP_i Sex_i + PC_{1i} + \cdots + PC_{10i} + \varepsilon_i \quad (5.1)$$

Where IPF_i is the log of odds of having IPF for subject i with SNP_i risk alleles, Sex_i and PC_{1i} to PC_{10i}, where $PC_1$to PC_{10} are the genetic principal components. $SNP_i Sex_i$ is the sex-interaction term. β_j is the log(OR) for a 1 unit increase in the j^{th} covariate. Within the analysis, the sex coding was female=0 and male=1. Therefore, for each SNP, the risk, or log(OR) for a female is given by β_1, with the log(OR) for a male given by $\beta_1 + \beta_2 + \beta_3$. Both β_1 and β_3 increment for each copy of the effect allele, whilst β_2 is constant within a SNP.

For each SNP in each of the individual studies the effect allele and results for both the SNP term (β_1) and the interaction term (β_3) were available. Variants included had gone through the same QC as the original GWAS (imputation quality >= 0.5, bi-allelic autosomal SNPs with HWE p-value <= $1x10^{-6}$). Additionally, only variants with a MAF >=0.01 and a call rate >=95% were included.

The 3 studies contain a total of 7 813 270 unique variants and the overlap between the input studies is shown in **Figure 5.1**.

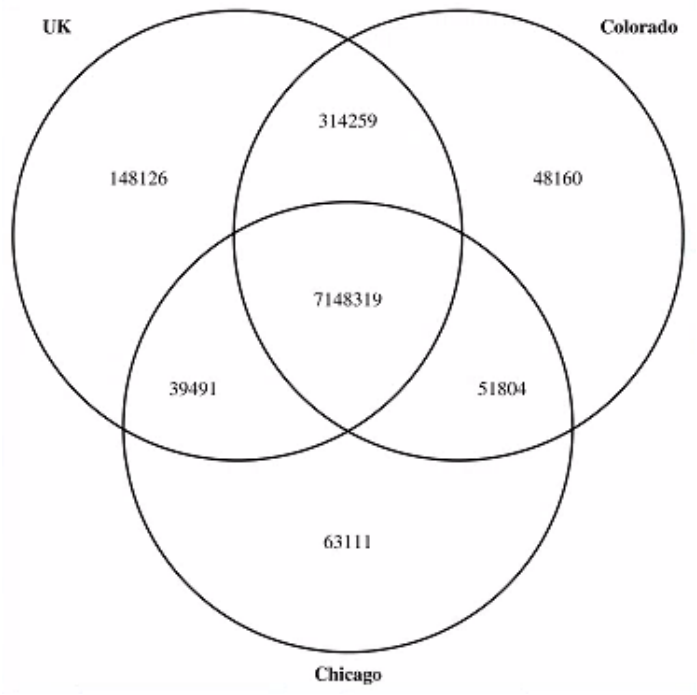


Figure 5.1: Venn diagram of variant overlap between input studies

5.1.2 GWA Meta-analysis

The 3 studies were meta-analysed using PLINK1.9 using a fixed-effect, inverse-variance weighted meta-analysis to give one effect size per SNP. Fixed effects were chosen as the assumption is that each SNP has a true underlying effect size, θ, that does not change by study (with covariates adjusted for), with any residual variation due to sampling error.

$$Y_i = \theta + e_i \tag{5.2}$$

The effect estimate from all 3 studies, $\hat{\theta}$, is:

$$\hat{\theta} = \frac{\sum w_i Y_i}{\sum w_i} \tag{5.3}$$

Where Y_i is the SNP effect estimate in study i and w_i is the weight given to each study. The weighting in the fixed effect model is given by 1 over the variance of the effect estimate.

$$w_i = \frac{1}{var(Y_i)} \tag{5.4}$$

Larger studies with smaller standard errors are given more weight than smaller studies. The variance of θ is 1 over the sum of all weights.

$$var(\theta) = \frac{1}{\sum w_i} \tag{5.5}$$

The fixed effect meta-analysis ensures that the method is consistent with that used for the sex-agnostic base data GWAS to allow direct comparison of results.

First the sex interaction term effects from β_3 in equation (*5.1*) were meta-analysed and the results explored. Variants that were present in at least 2/3 studies (7,553,873 in total) were kept for analysis.

Then, the coefficient for the main SNP association β_1, which in the model gives the effect for a female, was meta-analysed as described above. The data for this were raw PLINK output format, so an extra step was performed before meta-analysis to ensure that the effect allele (and corresponding effect directions) were harmonised across studies and consistent with the meta-analysis of the β_3 term.

5.1.3 PRS scoring Sex Interaction

As the p-values do not correspond to the overall SNP association, the PRS optimisation was performed as before with a low-resolution score, followed by high-resolution scoring.

First the meta-analysed sex interaction term (β_3) p-values + beta coefficients were used. This was done to explore whether a PRS score made up of variants with a significant sex interaction corresponds to a score with differential discriminative ability between the sexes.

5.1.4 Sex-interaction effect direction to inform SNP selection

Next the sex interaction term direction was used to inform the selection of SNPs for the creation of a PRS score based on the sex-agnostic effects (from the base data and effect sizes in chapter 4). If the genetic architecture differs between the sexes we may expect more extreme divergence of the PRS scores when applied to males and females if we condition on variants with a differential sex effect.

First a set of male specific SNPs was created by taking SNPs with a sex interaction term OR >1. From equation (*5.1*), this creates a set of SNPs where the differential effect between cases and controls is larger in males. The target data were restricted to just these male specific SNPS. PRS scoring was then performed. As in Chapter 4 the PRS was first optimised for a sex difference and the distribution of AUCs across a range of p-value thresholds was plotted. Then, the PRS that produces the highest AUC within each sex was extracted by splitting the target data and optimising for each sex independently. The performance of the best scores with each sex was compared to the

15 SNP GRS. In all this produced 3 separate analyses of the PRS thresholds for the Male specific SNP subset.

The above was repeated for a set of female specific SNPs (with a sex interaction OR < 1).

5.1.5 Sex-specific terms from interaction meta-analysis

The estimates for the SNP effects for females (β_1) from the sex interaction meta-analysis was considered. The results were plotted and explored to identify any variants of interest that had not been picked up in the sex-agnostic GWAS. These were defined as top SNPs that met the genome-wide suggestive threshold ($p < 1x10^{-5}$) using the female specific SNP effect estimate, with a MAF of > 5%, that had $p > 1x10^{-5}$ in the sex-agnostic GWAS.

Then the analyses as described in chapter 4 (section 4.1.3 to 4.1.5) were repeated using the female specific effect sizes as the base data. Low-resolution thresholds to identify the range for high-resolution scoring followed by high-resolution scoring. This was followed by 1) assessment of sex-differences in AUCs across the range of high-resolution p-value thresholds, and 2) optimising the PRS score by sex (4.1.5).

Finally, using the effect sizes from the main GWAS, variants of interest from the female specific results were added to the 12 SNP GRS using both the sex-agnostic and the female specific effect estimates, to determine if adding the female specific variants improved predictive accuracy in females.

5.2 Sex interaction term Results

5.2.1 Meta-analysis sex interaction term

No individual variants reached genome-wide significance (p-value threshold $5x10^{-8}$) in their sex interaction term (**Figure 5.2**). The *MUC5B* polymorphism on chromosome 11, which had a huge effect on the GRS AUC (and did not exhibit a sex difference in our earlier analysis) correspondingly has a sex interaction term with a high p-value (0.468) and low effect size (OR 1.083). In fact, none of the known risk variants for IPF stand out in **Figure 5.2**, agreeing with the lack of a sex difference seen in the GRS.

The QQ plot of the meta-analysed sex interaction term (**Figure 5.3**) shows an observed distribution reasonably close to the expected. This is consistent with the Manhattan plot (**Figure 5.2**) in that there are no variants with genome-wide association (or none that we have the power to detect). Additionally, overall population structure appears to be accounted for with the principal components, with further genomic control not indicated.

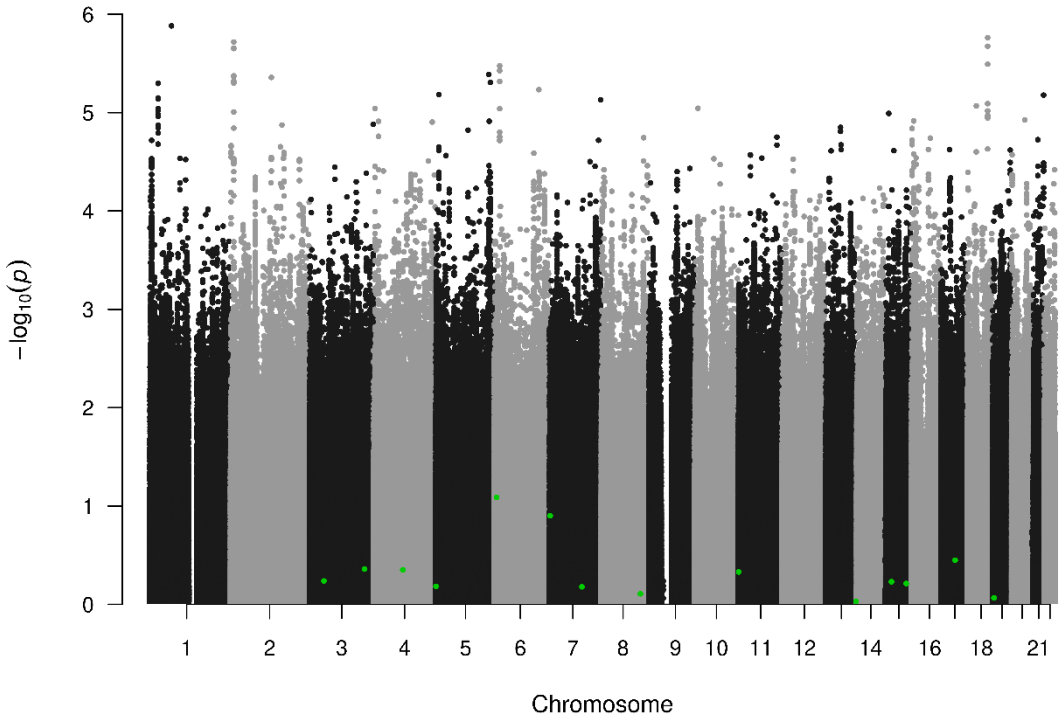


Figure 5.2: Manhattan plot of sex interaction p-values. Each dot represents a variant. Position is indicated along the x-axis, with the association of each variant given on the y-axis where higher -log10(p) indicates stronger association. Green dots represent SNPs used in the initial 15 SNP GRS (*SPDL1* not included as MAF <1%).

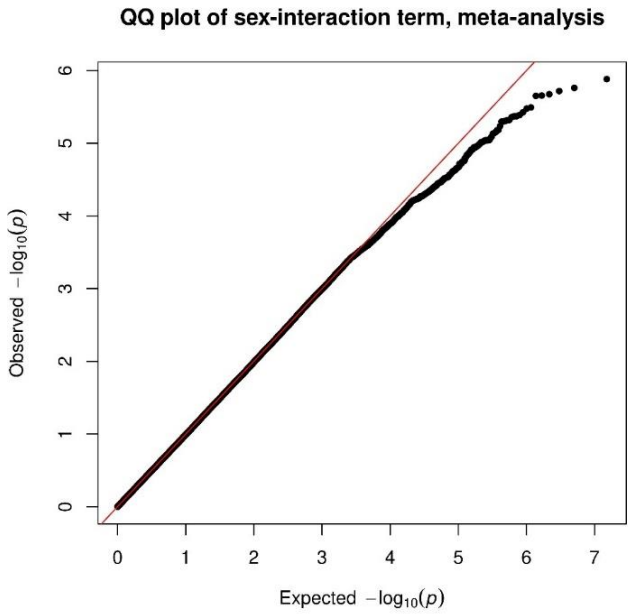


Figure 5.3: QQ plot of meta-analysis, sex-interaction term

5.2.2 PRS Scoring and Performance sex-interaction term and effects

Selected individual PRS p-value thresholds using the sex-interaction term are shown in **Figure 5.4**. None of the scores produced are significantly associated with IPF phenotype, with the most highly associated score at a threshold of 0.4 at p=0.113. The amount of variance explained as given by the liability R^2 is also very low.

The summary results from the default thresholds are shown in **Table 5.1**.

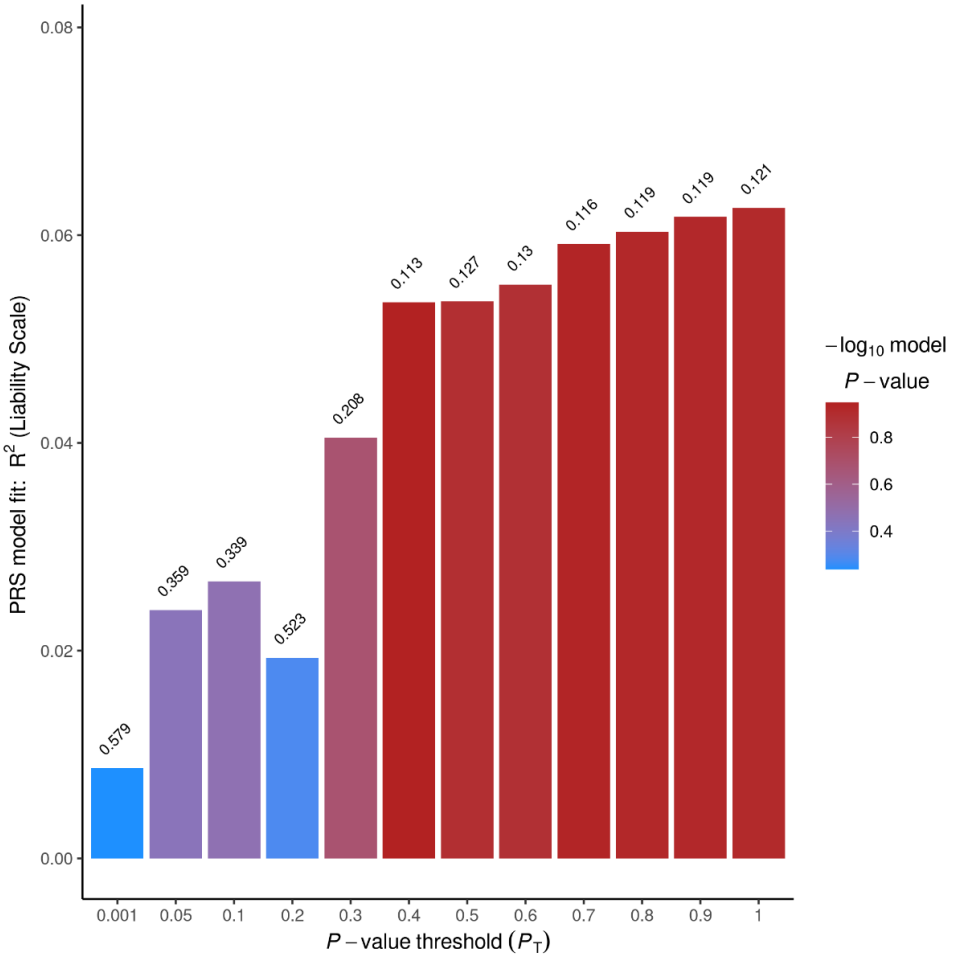


Figure 5.4: PRS using sex-interaction term, model fit at selected low-resolution thresholds

	ROC AUC			association
Threshold	male (95% CI)	female (95%CI)	Pval[1]	Score pval[2]
0.001	0.518 (0.494, 0.543)	0.527 (0.485, 0.569)	0.729	0.579
0.05	0.521 (0.497, 0.546)	0.541 (0.501, 0.582)	0.404	0.359
0.1	0.519 (0.495, 0.543)	0.540 (0.500, 0.580)	0.388	0.339
0.2	0.510 (0.486, 0.534)	0.555 (0.514, 0.596)	0.063	0.523
0.3	0.516 (0.491, 0.540)	0.555 (0.514, 0.596)	0.108	0.208
0.4	0.517 (0.493, 0.541)	0.559 (0.518, 0.599)	0.082	0.112
0.5	0.517 (0.493, 0.541)	0.558 (0.517, 0.599)	0.087	0.127

Table 5.1: Summary of Results from initial default PRS thresholds.

[1]p-value from de Long's test of 2 uncorrelated ROC curves

[2]Score p-value as generated by PRSice, from logistic regression including score + covariates

As there was no clear area of interest, high-resolution scoring was performed from 0.0001 to 1 in increments of 0.0001. The relative discriminative ability of the score does diverge by sex as more SNPs are added (**Figure 5.5**). However, the confidence interval for the male AUC contains 0.5 at all thresholds and is therefore considered no better than chance. There is less separation between males and females than seen with the primary analysis and confidence intervals overlap throughout.

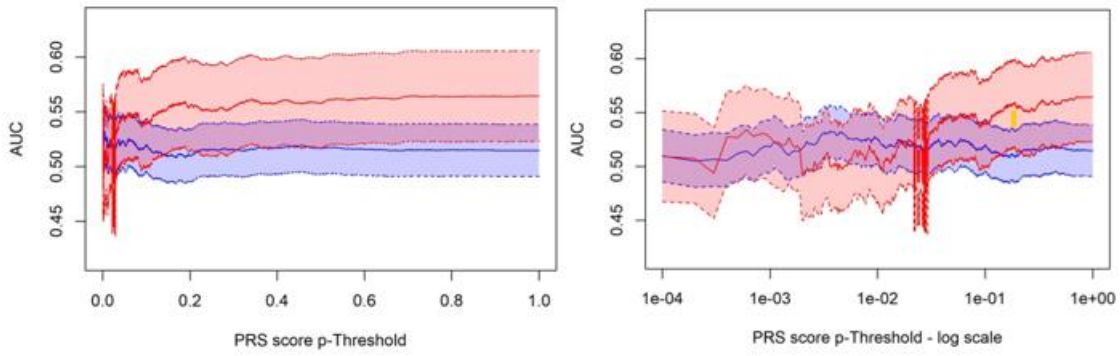


Figure 5.5: AUC by sex across p-value thresholds for PRS based on the sex interaction effect size. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI.

5.2.3 Using Effect direction to inform SNP selection

Within the sex interaction analysis there were 3,710,485 Male SNPs (OR > 1 in sex-interaction term), and 3,840,584 Female SNPs (OR < 1 in sex-interaction term). 2804 SNPs had an OR of exactly 1 and these were excluded from this analysis. Note that the *MUC5B* variant with an OR of 1.083 is considered a Male SNP here. After clumping and the removal of ambiguous SNPs, 235,933 Male SNPs and 243,289 Female SNPs remained for polygenic scoring using the sex-agnostic effect sizes.

After low-resolution exploratory scoring, high-resolution scoring was performed in the pT range 0.00005 to 0.3, this is wider than the range in Chapter 4. As before, additional scores with very stringent pTs were also calculated to help identify the threshold with the optimum discriminative ability.

With the Male SNP subset, males have a higher AUC at stringent pTs, but the male CI is completely encompassed by the female CI (**Figure 5.6**). The 2 sexes show the same (decreasing) performance until about pT=0.005. From then some divergence is seen between the sexes. This overall pattern very similar to the PRS performance on all

SNPs, though the divergence is less, with the biggest difference between the AUCs seen at pT 0.2968 (**Table 5.2**).

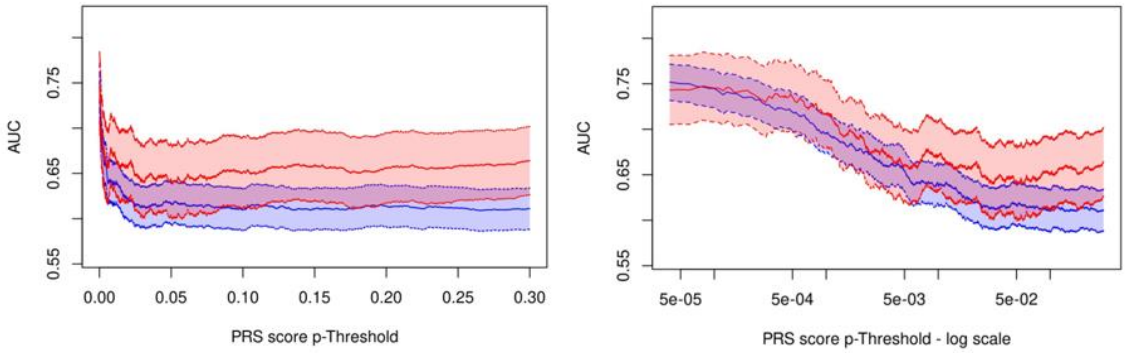


Figure 5.6: PRS Score using Male SNPs (OR > 1 in sex-interaction term), effect sizes from standard GWA. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI. Linear x-axis (right); Log x-axis (left).

Overall, the predictive accuracy in males is better compared to the main analysis with all SNPs. To illustrate, a pT of 0.1012 gives AUCs of 0.611 (0.588, 0.634) and 0.652 (0.613, 0.691) for males and females respectively. In the main analysis this threshold produced a lower male AUC and higher female AUC (0.593 and 0.659, **Table 4.2**).

For the subset of Female SNPs, the predictive accuracy is higher in females throughout (**Figure 5.7**). At low pTs the performance is noticeably lower than for the Male SNPs, but the *MUC5B* variant is not included in the female SNP set. The performance pattern again is similar to previous analyses, though with more sex divergence than seen for the Male SNPs with the biggest divergence at pT 0.0993 (**Table 5.2**). Here a pT of 0.1012 gives AUCs of 0.612 (0.588, 0.635) for males and 0.678 (0.643, 0.714) for females. This shows that the performance for males at this threshold is similar whether Male SNPs or Female SNP selections are used, however, for females, the Female SNPs do provide relatively better predictive accuracy at these less stringent pTs.

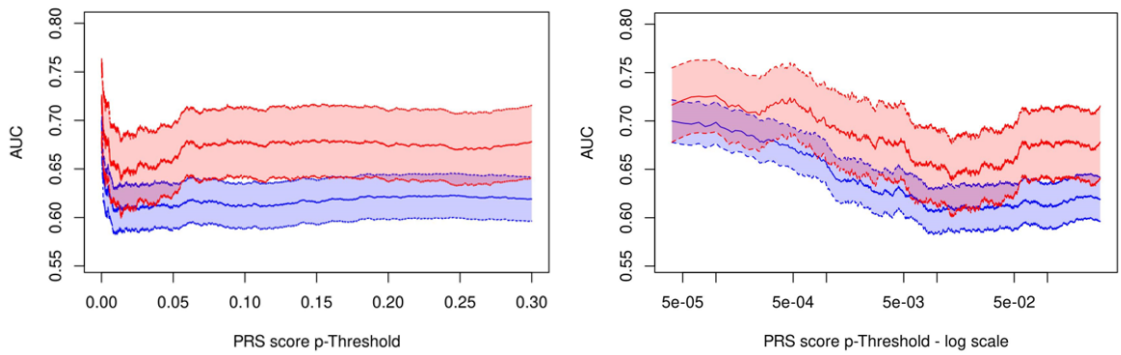


Figure 5.7: PRS Score using Female SNPs (OR < 1 in sex-interaction term), effect sizes from standard GWA. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI. Linear x-axis (right); Log x-axis (left).

5.2.4 Partitioned SNPs and the most predictive score by sex

When restricting the score to Male SNPs (OR > 1), the threshold that produced the most predictive score in males includes 44 SNPs at a pT of $5x10^{-6}$ while the most predictive score in females includes 53 SNPs at a pT of $1x10^{-5}$ (**Table 5.2**). However, there were no significant differences in performance in either sex between these 2 thresholds, with a p-value of 0.738 in males, and 0.257 in females. When compared to the 12 SNP GRS, the GRS outperformed both sex-specific scores, with nominal significance in males (p=0.048).

At their most predictive, the scores based on Female SNPs (OR < 1) had worse discrimination in both sexes compared to the score based on Male SNPs (**Table 5.2**).

SNP group	Threshold	ROC AUC male (95% CI)	ROC AUC female (95%CI)	Pval[1]	association pval[2] male	association pval[2] female
Biggest sex difference						
OR > 1	0.2968	0.610 (0.588, 0.633)	0.664 (0.626, 0.701)	0.0176	$3.40x10^{-37}$	
OR < 1	0.0993	0.611 (0.588, 0.634)	0.679 (0.643, 0.715)	0.0017	$8.20x10^{-35}$	
Most predictive PRS						
OR > 1	male:$5x10^{-6}$	**0.754** (0.734, 0.775)	0.743 (0.705, 0.780)	0.5874	$1.44x10^{-95}$	$4.99x10^{-33}$
OR > 1	female:$1x10^{-5}$	0.754 (0.734, 0.774)	**0.747** (0.710, 0.784)	0.7468	$6.86x10^{-94}$	$1.02x10^{-33}$
OR < 1	male:$1x10^{-4}$	**0.699** (0.677, 0.720)	0.726 (0.689, 0.763)	0.2111	$1.54x10^{-61}$	$3.34x10^{-27}$
OR < 1	female:$1x10^{-4}$	0.699 (0.677, 0.720)	**0.726** (0.689, 0.763)	0.2111	$1.54x10^{-61}$	$3.34x10^{-27}$

Table 5.2: Performance of PRS scores, using sex interaction effect direction

[1]p-value from de Long's test of 2 ROC curves

[2]Score p-value as generated by PRSice, from logistic regression including score + covariates. Produced by PRS scoring by each sex separately

5.3 Female specific SNP association results

5.3.1 Meta-analysis female specific SNP term

The results from the meta-analysis of the β_1 term are shown in **Figure 5.8**. This gives the individual SNP associations for females. It was expected that not all the known genetic risk variants for IPF found in the main GWAS would reach genome-wide significance.

The QQ plot in **Figure 5.9** indicates true association too, with the divergence from the expected line from around $-\log_{10}(p)$ 5. Some of that will be due to the very strong association of *MUC5B,* which still has an OR of 4.53, p= $2.53x10^{-67}$ in females.

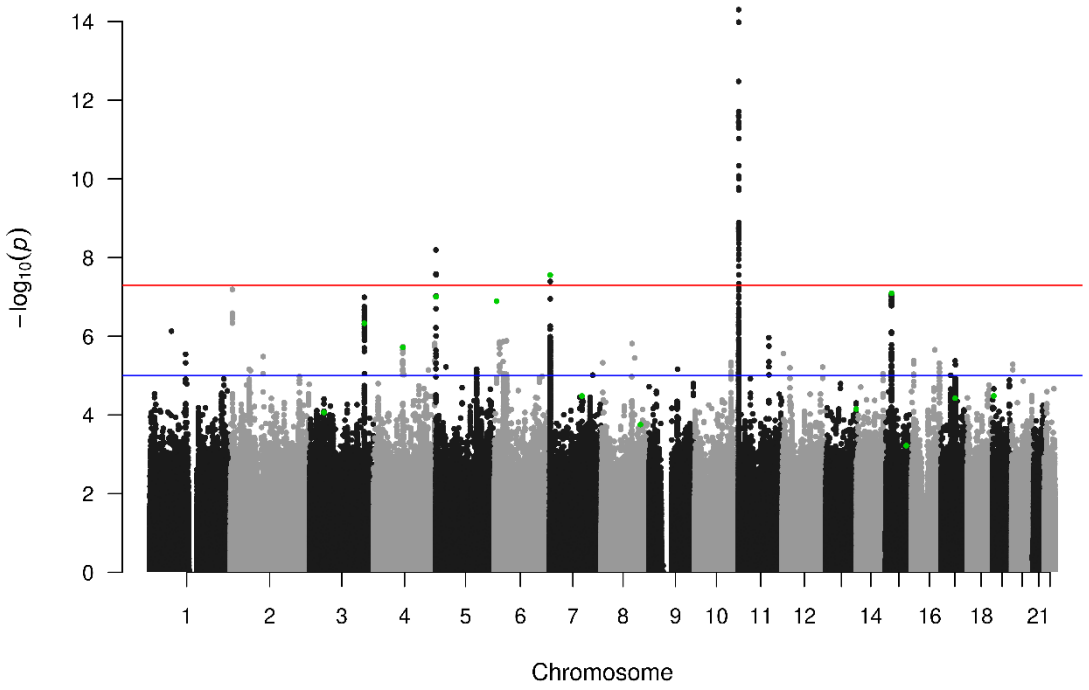


Figure 5.8: Manhattan Plot of β_1 SNP term (for females) p-values. Each dot represents a variant. Position is indicated along the x-axis, with the association of each variant given on the y-axis where higher $-\log_{10}(p)$ indicates stronger association. Red horizontal line at $5x10^{-8}$ is the genome-wide significance line, blue line at $1x10^{-5}$ the genome-wide suggestive line. Green dots represent SNPs used in the initial 15 SNP GRS (*SPDL1* not included as MAF <1%). Y axis truncated at 15 (*MUC5B* variant not shown).

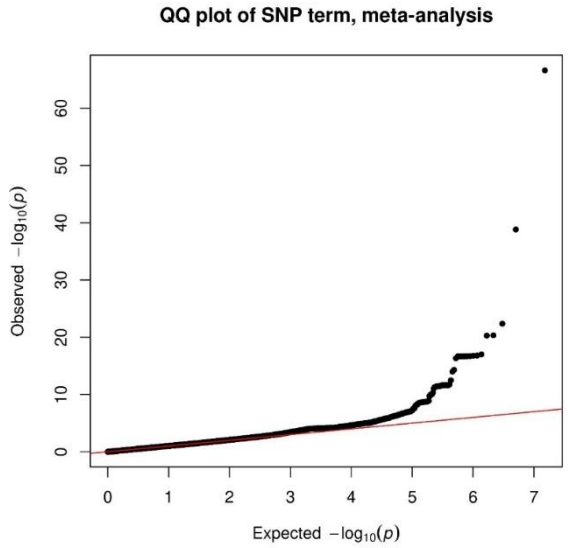


Figure 5.9: QQ plot of meta-analysis, β_1 (SNP) term for females

Chrom	Position	SNP	Gene	MA	MAF	effect size	p-value
1	67280355	rs2454321*	*DNAI4*	C	0.120	0.596	$7.39x10^{-7}$
1	111015838	rs11102122*		G	0.420	0.760	$2.86x10^{-6}$
2	5923038	rs76044167		A	0.019	2.551	$6.42x10^{-8}$
2	57874707	rs2612311*		T	0.396	1.289	$6.93x10^{-6}$
2	100647704	rs116177376	*AFF3*	C	0.018	2.623	$3.27x10^{-6}$
3	169486508	rs35446936	*ACTRT3*	A	0.262	1.392	$1.02x10^{-7}$
4	89883979	rs7671167	***FAM13A***	T	0.479	0.765	$1.83x10^{-6}$
4	168758989	rs75226502		A	0.015	2.472	$7.15x10^{-6}$
4	188196933	rs73015913*		T	0.129	1.458	$1.52x10^{-6}$
5	1285974	rs7705526	***TERT***	A	0.308	0.674	$6.38x10^{-9}$
5	32091868	rs788368*	*PDZD2*	T	0.365	1.293	$6.02x10^{-6}$
5	126816667	rs57256097*		A	0.140	1.411	$8.84x10^{-6}$
6	7563232	rs2076295	*DSP*	G	0.472	1.344	$1.29x10^{-7}$
6	17015381	rs11754838*		A	0.147	0.657	$1.39x10^{-6}$
6	28728329	rs1233597*		T	0.274	1.342	$1.36x10^{-6}$
6	37976827	rs17648699*	*ZFAND3*	C	0.099	1.521	$1.31x10^{-6}$
7	1868872	rs138576161	*MAD1L1*	T	0.019	2.516	$5.43x10^{-6}$
7	1909479	rs12699415	*MAD1L1*	A	0.422	1.372	$2.77x10^{-8}$
7	132924778	rs75334297*		A	0.055	1.756	$9.67x10^{-6}$
8	95125274	rs116894624		A	0.021	2.355	$1.53x10^{-6}$
8	102845921	rs148347297	*NCALD*	T	0.024	2.489	$3.55x10^{-6}$
9	88830169	rs62571526		G	0.009	3.133	$6.83x10^{-6}$
10	112538762	rs4147101*	*RBM20*	G	0.443	0.768	$4.60x10^{-6}$
10	134716909	rs59357541	*CFAP46*	T	0.030	1.939	$3.04x10^{-6}$
11	1241221	rs35705950	*MUC5B*	T	0.151	4.532	$2.53x10^{-67}$
11	94165937	rs187036511	*MRE11*	C	0.029	2.124	$1.78x10^{-6}$
12	4153562	rs117122179		A	0.009	3.460	$2.74x10^{-6}$
12	24375959	rs35125885*	*SOX5*	T	0.357	1.305	$6.31x10^{-6}$
12	125254373	rs4765607*		C	0.168	0.680	$6.05x10^{-6}$
14	101580779	rs12879668*		G	0.474	1.309	$9.05x10^{-6}$
15	40720542	rs59424629	*IVD*	G	0.453	0.741	$8.04x10^{-8}$
16	6739091	rs2110351*	*RBFOX1*	A	0.171	1.379	$4.13x10^{-6}$
16	71757230	rs147174874	*PHLPP2*	C	0.020	2.199	$2.20x10^{-6}$
16	86000964	rs9934574		T	0.245	1.337	$4.82x10^{-6}$
17	30476819	rs144949591	*RHOT1*	A	0.020	2.208	$9.77x10^{-6}$
17	43946370	rs242922	*MAPT-AS1*	A	0.418	1.307	$4.21x10^{-6}$
20	3662736	rs6084435*	*ADAM33*	G	0.082	1.555	$5.16x10^{-6}$

Table 5.3: Possible hits for females from sex interaction GWAS, $p < 1x10^{-5}$

MA=Minor allele; MAF= Minor allele frequency (average of the frequencies in the contributing studies). Effect size (odds ratios) and p-values from meta- sex interaction GWAS. Rows in grey are known GWAS associations, brown are not female specific. Rows in black are female specific (i.e., main GWAS p-value < $1x10^{-5}$); underlined indicates female specific SNPs with MAF> 5%.

The top SNPs above the genome-wide suggestive line ($p < 1x10^{-5}$) were extracted and are shown in **Table 5.3**. This produced 16 female specific SNPs with a MAF > 5%. Of these, 7 may be associated with specific genes.

5.3.2 PRS Scoring and Performance female specific SNP term

Unlike the sex-interaction term, PRS scoring on the β_1 SNP term, did produce PRS scores that were associated with phenotype (**Figure 5.10**), though less strongly compared to the PRS associations seen with sex-agnostic weightings. The results shown are for the performance of the overall PRS (males and females combined).

When considering the relative performance of these PRS scores by sex (**Figure 5.11**), little difference is seen. The AUC across the whole distribution is also lower compared to the main analysis, stabilising at around 0.55 compared to 0.60 for males and 0.65 for females. The best predictive accuracy is seen in a pT of $5x10^{-7}$ (**Table 5.4**) for both sexes. At this threshold the female AUC is slightly higher than the male AUC, but not significantly so (p=0.825). The AUC is also lower than the maximum achieved in the main analysis (males 0.753, females 0.747, **Table 4.3**).

Some small sex differences are seen between pT 0.001 and 0.05, though these do not approach significance when corrected for multiple testing, with the biggest performance difference at pT 0.0059 (p=0.012) (**Table 5.4**).

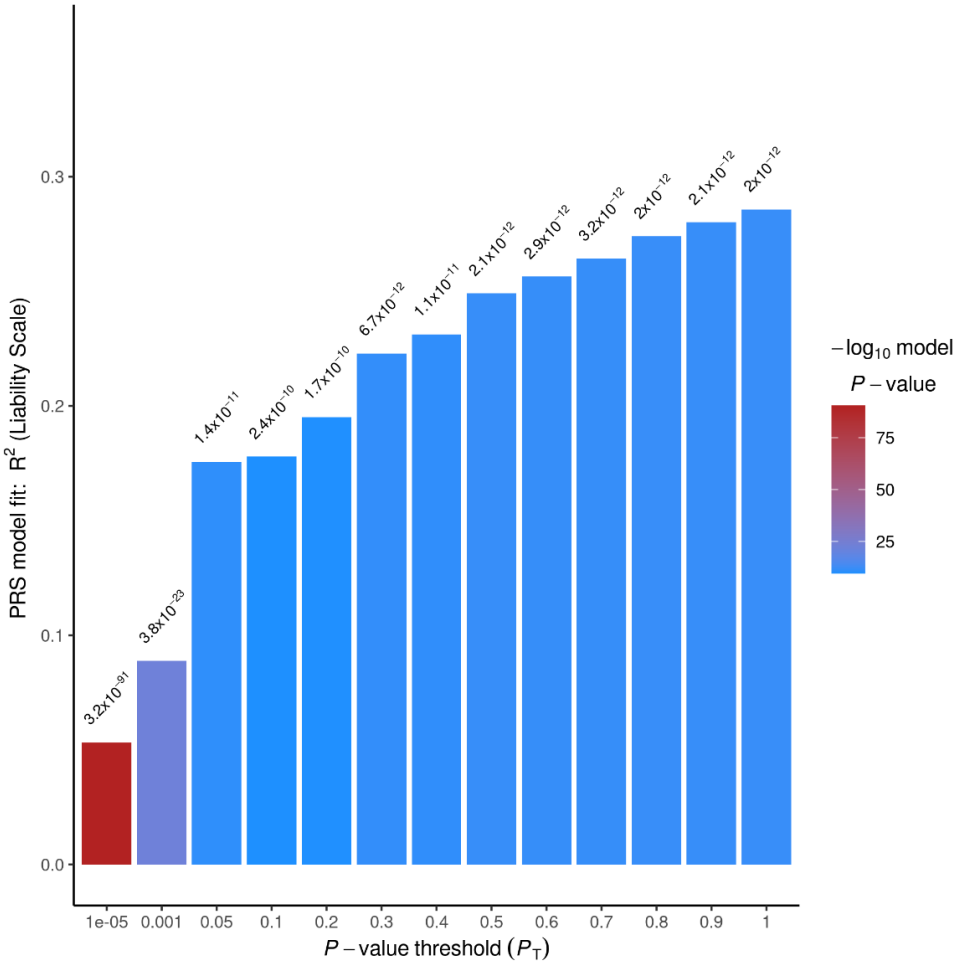


Figure 5.10: PRS association of β1 SNP term at selected pTs. R^2 on the liability scale assuming a population prevalence of 63 cases per 100,000

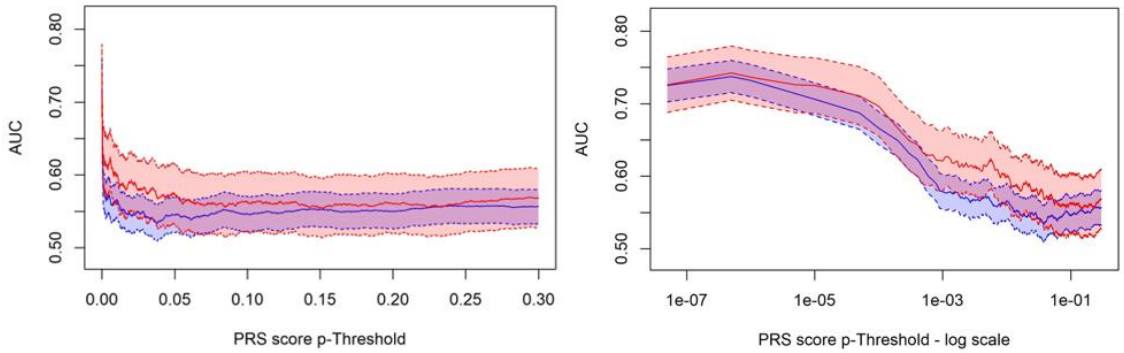


Figure 5.11: AUC from PRS Score using β1 SNP term from sex interaction analysis. Red area indicates female score and 95% CI; blue area indicates the male score and 95% CI. Linear x-axis (right); Log x-axis (left).

Threshold		N SNPs	ROC AUC male (95% CI)	ROC AUC female (95%CI)	ROC AUC Pval[1]	Score P value[2]
Biggest Sex difference						
0.0059		7134	0.562 (0.538, 0.587)	0.622 (0.582, 0.661)	0.012	$9.82x10^{-18}$
Best male	$5x10^{-7}$	15	0.737 (0.715, 0.759)	0.742 (0.705, 0.779)	0.825	$3.81x10^{-116}$
Best female	$5x10^{-7}$	15	0.737 (0.715, 0.759)	**0.742** (0.705, 0.779)	0.825	$3.81x10^{-116}$

Table 5.4: Relative Performance of best Male and Female thresholds

[1]p-value from de Long's test of 2 ROC curves

[2]Score p-value as generated by PRSice, from logistic regression including score + covariates.

5.3.3 GRS + female specific variants of interest

Addition of female hits with MAF > 5% added 16 SNPs to the 12 SNP GRS (14 after clumping). Addition of female hits of any MAF added a further 11 SNPs. As seen in **Table 5.5**, adding these female specific SNPs did not improve the performance of the expanded GRS in either sex relative to the 12 SNP GRS. This was true with both sex-agnostic and female specific effect estimates.

Score	N SNPs	association Score pval[1]	ROC AUC male (95% CI)	ROC AUC female (95%CI)
Sex-agnostic effect estimates				
12 SNP GRS	12	9.11×10^{-156}	0.783 (0.761, 0.805)	0.778 (0.739, 0.817)
GRS + female SNPs MAF >5%	26	9.02×10^{-151}	0.778 (0.759, 0.798)	0.765 (0.737, 0.803)
GRS + female SNPs	37	5.48×10^{-144}	0.773 (0.753, 0.793)	0.757 (0.720, 0.796)
Female specific effect estimates[3]				
GRS	11	$1.80x10^{-149}$	0.780 (0.760, 0.799)	0.757 (0.719, 0.796)
GRS + female SNPs MAF >5%	25	$3.47x10^{-113}$	0.732 (0.711, 0.754)	0.742 (0.703, 0.782)
GRS + female SNPs	36	$7.22x10^{-92}$	0.713 (0.691, 0.734)	0.719 (0.678, 0.759)

Table 5.5: Performance of GRS + female hits, compared to GRS using sex-agnostic and female specific effect sizes

[1]p-value of score from model in equation (3.2)

[3]*SPDL1* variant missing from female effect estimates due to MAF < 1%

5.4 Discussion

5.4.1 Sex-interaction term assessment

Summing variants based on their sex-interaction coefficient does not magnify any sex differences, producing scores with poor discriminative ability. No individual score reached a statistically significant sex difference in AUC. This is due to summation of relative differences that do not correspond to phenotype. At each threshold the score contains both variants with a larger effect in males, and variants with a larger effect in females. The direction of effect is also problematic in this analysis. Take for example a SNP that confers risk, but the increase in risk is greater for females than males. This produces a negative effect weighting as input to the PRS. Conversely, a SNP that is protective overall, but with less protection conferred to males will have a positive weighting.

The value in this analysis is in showing that by considering the p-values and effect sizes from the sex-interaction term as input to a PRS is not logical and it tells us very little about the sex differences within the genetics of IPF. The expectation of a null result when using the sex interaction term could be interesting to verify with a simulation study.

5.4.2 Sex-interaction term direction to inform SNP selection

Using the direction of the sex-interaction term to partition the sex-agnostic GWAS base data into Male SNPs and Female SNPs did not confer any advantages. Though the female PRS score became more predictive relative to the main analysis at pTs > 0.05, the resulting sex difference did not meet the 0.001 significance threshold (p=0.0017).. The score with the best predictive accuracy in males, using the male SNPs was

comparable to the most predictive PRS using all SNPs, with AUCs of 0.754 (0.734, 0.775) and 0.753 (0.730, 0.776) respectively. The narrower confidence interval around the male AUC when Male SNPs only are considered implies that no information is lost. When using Male SNPs, the male performance is better than female performance and vice versa with the Female SNPs, though differences are not statistically significant. However, SNPs with an interaction effect may have a sex-agnostic effect size that is attenuated towards the null in the sex-agnostic GWAS, especially where sex-specific effects are in opposite directions. These are not accounted for in this analysis.

5.4.3 Female specific effect estimates

With such a small sample size (806 female cases), there are likely to be false positives in the SNP associations, and we should be cautious when interpreting the results from the female specific effects from the sex interaction model as a 'discovery GWAS'. Despite this, findings may help to increase our understanding of the disease or identify druggable targets for further consideration.

The *ADAM33* (ADAM Metallopeptidase Domain 33) gene is a type 1 transmembrane protein involved in cell-cell and cell-matrix interactions. That has been associated with asthma and bronchial hyper-responsiveness[101]. The protein it encodes is part of the matrix metalloproteinase (MMP) group. Several proteins in this group (MMP1 and MMP7) are highly expressed in IPF lungs and are potentially useful biomarkers in IPF, both for differentiation from related lung diseases and prognosis[102]. *ADAM33* has been associated with IPF under a recessive model in Chinese and Japanese patients in a previous study[103] though to date this finding has not been replicated or associated with IPF in European subjects. In addition, the area around rs6084435 exhibits high LD

(see the region plot in APPENDIX E) and the causal variant for the association may not be in *ADAM33*. It is possible that the *ADAM33* gene is of sex-specific functional significance in the development of IPF, but further work will be needed to explore this.

Several other genes (*ZFAND3, SOX5, RBFOX1*) appear to be involved with translation and transcription, but these have not been linked specifically with any lung diseases.

Of the variants with a MAF < 5%, *MRE11* is involved in the repair and maintenance of DNA repair and telomere maintenance[104].

The PRS that used the female specific effect size estimates and p-values did not show any significant sex differences or performance advantage over the main analysis. The estimates used here will have much greater uncertainty than even the sex-agnostic base data. The results at very stringent p-value thresholds, were comparable to the main analysis as the *MUC5B* and several other known IPF risk variants were included. These are also variants where the sex-interaction effect is low, so the SNP effect is almost completely captured by the β_1 term.

Whilst in theory, using results from a sex interaction GWAS should allow for the investigation of sex differences in traits, a much bigger sample size is needed for the resulting score to be useful.

5.5 Summary

In this chapter sex-specific estimates from a sex-interaction GWAS were considered. First results from the 3 separate IPF studies were meta-analysed. After establishing that a PRS consisting of the cumulative sex interaction effect sizes and p-values was

not helpful in exploring sex differences, the sex interaction effect direction was used to inform SNP selection from the sex-agnostic base GWAS. This did not highlight any additional sex differences compared to the main analysis.

The female specific effect estimates from the sex-interaction analysis were considered. Exploring the meta-analysis results highlighted several SNPs with a female specific effect that could be considered for further study including a variant in *ADAM33*. Despite this, PRS scores from the female estimates performed worse than the scores in the main analysis. Adding the female specific SNPs of interest to the GRS did not improve its predictive accuracy in females in our target data.

6 DISCUSSION

In this final chapter, the aims of the project are restated and major findings summarised. These will be put into context and avenues for future further research are discussed.

6.1 Overall Findings

The main objectives of this project were to use PRS to explore

1) The existence of any sex differences in the genetic architecture of IPF.

2) Whether taking sex into account can help to improve prediction accuracy of a PRS in a clinically meaningful way.

The initial analysis of the 15 SNP GRS using sex-agnostic effect sizes showed that the combining effects from the 15 known IPF risk variants produced a score that was highly associated with IPF (OR 2.67; $p=1.49 \times 10^{-160}$) with a moderately high discriminative ability between cases and controls (AUC 0.787 CI: 0.768 to 0.806). Stratifying the score by sex did not improve predictive accuracy (males 0.791 CI: 0.769 to 0.814); females 0.773 CI: 0.733 to 0.812; p = 0.385). The clinical utility of the score was considered and, despite the moderately high AUC, the positive predictive value at population level was low.

To expand the GRS to PRS scoring with many p-value thresholds, PRSice v1.25 was used. A technical replication was performed to ensure comparable results.

Some sex differences were seen when performing PRS scoring using the sex-agnostic weightings with a maximum difference at around pT=0.1. Restricting the target data to variants with MAF > 5%, resulted in a significant sex difference at pT 0.1 - 0.15.

Optimising for predictive accuracy by sex did not produce improved scores compared to the GRS (males p value threshold $1x10^{-6}$, AUC 0.753 CI: 0.730 to 0.776; females p value threshold $1x10^{-5}$, AUC 0.747 CI: 0.707 to 0.788). In other words, these analyses yielded no improvements over the GRS.

Using summary statistics from the sex interaction meta-analysis to inform SNP selection of sex-specific subsets did not improve scores either, with the 'Male SNP' (sex-interaction OR > 1) selection giving comparable performance to the main analysis, with slightly narrower confidence intervals (males p value threshold $5x10^{-6}$, AUC 0.754 CI: 0.734 to 0.775; females p value threshold $1x10^{-5}$, AUC 0.747 CI: 0.710 to 0.784). The 'Female SNP' selection did not perform as well.

Using meta-analysing the SNP term from the sex interaction analysis to provide female specific effect estimates highlighted 16 variants of interest identified including a variant in *ADAM33,* a gene which has been previously implicated in IPF. When these were added to the GRS they did not improve the predictive accuracy with either sex-agnostic or female-specific effect estimates.

6.2 Limitations and Further Research

6.2.1 GWAS effect estimates

For accurate prediction of risk, the base data sample size needs to be as large as possible, and with 2668 cases (of which 1811 are male and 806 female), there is likely

to be high uncertainty in the estimates of the individual SNP associations sampling error is increased, bringing us back to the limited power in the current IPF GWAS. Restricting the PRS scoring to common variants (MAF > 5%) as done in the Chapter 4 sensitivity analysis may help with some of these limitations as the effect estimates of rare variants will be more affected by the lack of power and are more likely to have genotyping issues[105]. However, in this analysis the data was restricted based on the MAF in the target data when it was more appropriate to restrict the base data instead. With this, some independent signals may be lost with corresponding decrease in predictive ability. In terms of PRS optimisation, we are likely to be overfitting to our target data across analyses in this project – modelling the target data too closely at the expense of generalisability[77] and to confirm or validate these results a third sample will be needed.

Extremely strongly associated PRS scores were produced at low p-value thresholds with the overall best score coming from the 15 SNP GRS with just the known IPF risk variants. From previous work by Dudbridge[82], it is known that if a trait has a low number of associated variants (i.e. a low proportion of SNPs that contribute to disease risk), then the p-value threshold that optimises predictive accuracy is low, irrespective of the proportion of the heritability of liability explained by the marker panel. This is in line with findings by Allen (PhD thesis) who investigated the polygenicity of IPF using PRS analyses and concluded that IPF is a trait with a small number of variants that have a relatively large effect on susceptibility[91]. The overall effect in this data is that the contribution of *MUC5B* on the results appears to be diluted as more SNPs are added.

6.2.2 Sex-Specific GWAS effect estimates

Another obvious limitation is the lack of base data effect estimates from sex-specific GWAS. A recent study by Bernabeu et al.[62] that investigated sex differences in 530 traits of which 446 were binary, using the UK Biobank did use sex-specific GWAS estimates with their sex-stratified polygenic scores. Similar methods were employed by Han et al. (pre-print)[106]. The use of sex-specific GWAS also allows for the calculation of sex-specific heritability. Sex can then be considered as a trait, and the assessment of genetic overlap between the sexes investigated through methods such as LD score regression[107,108]. Unfortunately, performing a sex stratified GWAS on the data available in this project, with just 806 female cases, and 1811 male cases is not likely to yield individual variant effect estimates with a precision to allow meaningful PRS scores to be created, especially in females and much larger sample sizes will be needed[78].

6.2.3 Bias in effect size estimates – Winner's curse

The Winner's curse states that the winning bid in an auction tends to be greater than the intrinsic value of that item[109], in genetics it means that effect sizes for associated variants as reported in GWAS studies show an upward bias in the discovery analysis[110].

This effect can be seen in our data. 3 new genome-wide associations in *KIF15,* *MAD1L1*, and *DEPTOR* were reported from the sex-agnostic GWAS. The effect sizes from the KIF15 and *DEPTOR* variants were larger, OR 1.58 v 1.48 and OR 0.82 v 0.87 respectively between discovery and replication[34]. The GWAS base data also contained genome-wide associations that did not replicate and where the winner's curse bias is even more extreme. For example, a variant in *HECTD2* had an OR of 7.82 (p 4.43×10^{-8}) in the base data, but in replication the OR was just 1.75 (p=0.155)[34].

True unbiased estimates can only be drawn from large population samples using a cohort design[111]. This is not feasible in rare diseases such as IPF. Methods to adjust effect sizes to create unbiased estimators[110] and PRS methods that attempt to correct for this bias have been developed[112], but these were not evaluated within this project.

Another feature of the upward bias in GWAS effect estimates is that in low powered studies, as the power to detect female specific effects was, the effect size in significant hits may have little bearing on the true effect size[113]. This means that the PRS on female only effects may have been of limited value and that our identification of variants of interest is exploratory at best.

6.2.4 PRS method used

In this project the C+T method for generating PRS scores was implemented using PRSice. One limitation of PRSice is the deletion of ambiguous SNPs, which potentially loses information where no proxies are available. When ambiguous SNPs were included the sex difference seen was reduced compared to the main analysis (Chapter 4), with higher male AUCs (little change in female AUCs). If removing the ambiguous SNPs removes information then the true sex-difference is likely to be less than seen in the main analysis. If excluding the ambiguous SNPs leads to loss of many independent signals (from low frequency functional variants without proxy) we may expect both male and female discriminative ability to improve, which did not occur. As a result, it seems more likely that results seen are due to an overall increase in number of SNPs included at each threshold.

Other PRS methods are available with differing approaches to deal with LD and effect size estimation (through shrinkage). Rather than clumping or pruning to create a set of partially independent SNPs, it is also possible to model the LD. This method keeps all SNPs within the score, but accounts for the LD between them, by modelling them jointly[114].

In the C+T method used, no changes are made to the effect size estimations of individual SNPs, but the p-value thresholding shrinks the effect sizes of all SNPs that are excluded from the score at a particular threshold to zero[77]. Instead, it is possible to apply shrinkage to all SNPs using penalised regression such as LASSO[115] or ridge regression and several variations have been proposed[116,117] and implemented in software such as lassosum[118]. There are also approaches that use a Bayesian framework to perform shrinkage via prior distributions. Software implementations include LDPred[114], PRS-CS[119] and JAMPred[120].

The optimal shrinkage factor depends on the underlying true effect size distribution and the mixture of null SNPs and those with a real effect size. As a result, as for p-value thresholding, the score must be optimised across a range of possible shrinkage parameter values. Though shrinkage estimation methods reduce the mean square error of the prediction, in practical terms, they do not improve the estimates over those produced by standard logistic regression when used for association testing and AUC[82,121] and the performance gains to date are minimal[122].

6.2.5 PRS Clinical utility

In Chapter 3 the clinical utility of the GRS was considered. Since this was the overall score with the best predictive accuracy in both sexes, its performance and clinical utility can be considered in the context of future work.

The score might already be useful in high prevalence contexts. For example, IPF is estimated to account for 17-37% of newly diagnosed ILD cases[16]. A large challenge in IPF diagnosis is excluding other types of ILD once a patient is known to have ILD. Considering the GRS in Chapter 3, and using those results to calculate PPVs, now for every 100 ILD patients, up to 80 have IPF (cut-off dependent). Further work is needed to ascertain if the genetic overlap between IPF and other ILDs is small enough for a score created using normal controls and IPF patients to be of benefit. If so, an addition of genetic information in the form of a GRS score to existing prediction models may improve their performance.

It has been shown that combining genetic components with clinical predictors can improve the performance of the resulting score in IPF prognosis[72]. This project did not explore the prediction accuracy when clinical predictors are combined with any of the polygenic scores, and this should be considered in future work. Even though we found no evidence that a sex-specific PRS should be used in IPF risk prediction, a clinical score made up of sex and the PRS is likely to have better discriminative ability than the PRS score alone.

6.3 Overall Conclusions

Within this project we have established that PRS scores can be used to investigate sex differences in genetic architecture of a rare complex trait such as IPF.

A GRS of just the known IPF risk variants had moderate predictive accuracy with an overall AUC of 0.787 (0.768, 0.806). None of the more complex analyses improved on this.

More work needs to be done with larger datasets and sex-specific GWAS before we can conclude whether there are sex differences within the genetics of IPF. However, at present, there is no evidence to recommend that sex-specific PRS scores be considered for risk prediction.

REFERENCES

1. Raghu, G. *et al.* An Official ATS/ERSARS/ALAT Statement: Idiopathic Pulmonary Fibrosis: Evidence-based Guidelines for Diagnosis and Management. *American Journal of Respiratory and Critical Care Medicine* **183**, 788-824 (2011).
2. Ley, B., Collard, H.R. & King, T.E., Jr. Clinical Course and Prediction of Survival in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **183**, 431-440 (2011).
3. Hutchinson, J., Fogarty, A., Hubbard, R. & McKeever, T. Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review. *European Respiratory Journal* **46**, 795-806 (2015).
4. Rudd, R.M., Prescott, R.J., Chalmers, J.C., Johnston, I.D.A. & Fibrosing Alveolitis Subcomm Res, C. British Thoracic Society Study on cryptogenic fibrosing alveolitis: response to treatment and survival. *Thorax* **62**, 62-66 (2007).
5. King, T.E. *et al.* Effect of interferon gamma-1b on survival in patients with idiopathic pulmonary fibrosis (INSPIRE): a multicentre, randomised, placebo-controlled trial. *Lancet* **374**, 222-228 (2009).
6. Hyldgaard, C., Moller, J. & Bendstrup, E. Changes in management of idiopathic pulmonary fibrosis: impact on disease severity and mortality. *European Clinical Respiratory Journal* **7**(2020).
7. Lederer, D.J. & Martinez, F.J. Idiopathic Pulmonary Fibrosis. *New England Journal of Medicine* **378**, 1811-1823 (2018).
8. Raghu, G. *et al.* Diagnosis of Idiopathic Pulmonary Fibrosis An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *American Journal of Respiratory and Critical Care Medicine* **198**, E44-E68 (2018).
9. Thickett, D. *et al.* Historical database cohort study addressing the clinical patterns prior to idiopathic pulmonary fibrosis (IPF) diagnosis in UK primary care. *Bmj Open* **10**(2020).
10. Kim, H.J., Perlman, D. & Tomic, R. Natural history of idiopathic pulmonary fibrosis. *Respiratory Medicine* **109**, 661-670 (2015).
11. Collard, H.R. *et al.* Acute Exacerbation of Idiopathic Pulmonary Fibrosis An International Working Group Report. *American Journal of Respiratory and Critical Care Medicine* **194**, 265-275 (2016).
12. Prasad, R., Gupta, N., Singh, A. & Gupta, P. Diagnosis of idiopathic pulmonary fibrosis: Current issues. *Intractable & Rare Diseases Research* **4**, 65-69 (2015).
13. Tepede, A. & Yogaratnam, D. Nintedanib for Idiopathic Pulmonary Fibrosis. *Journal of Pharmacy Practice* **32**, 199-206 (2019).
14. Ley, B. *et al.* Pirfenidone Reduces Respiratory-related Hospitalizations in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **196**, 756-761 (2017).
15. Gerstman, B.B. *Epidemiology Kept Simple : An Introduction to Traditional and Modern Epidemiology*, (John Wiley & Sons, Incorporated, Hoboken, UNITED KINGDOM, 2013).
16. Nalysnyk, L., Cid-Ruzafa, J., Rotella, P. & Esser, D. Incidence and prevalence of idiopathic pulmonary fibrosis: review of the literature. *European Respiratory Review* **21**, 355-361 (2012).
17. Marshall, D.C., Salciccioli, J.D., Shea, B.S. & Akuthota, P. Trends in mortality from idiopathic pulmonary fibrosis in the European Union: an observational

study of the WHO mortality database from 2001-2013. *European Respiratory Journal* **51**, 9 (2018).
18. Annesi-Maesano, I. *et al.* EPIDEMIOLOGY OF IDIOPATHIC PULMONARY FIBROSIS IN EUROPE - AN UPDATE. *Sarcoidosis Vasculitis and Diffuse Lung Diseases* **30**, 6-12 (2013).
19. Raghu, G. *et al.* Idiopathic pulmonary fibrosis in US Medicare beneficiaries aged 65 years and older: incidence, prevalence, and survival, 2001-11. *Lancet Respiratory Medicine* **2**, 566-572 (2014).
20. Moore, C. *et al.* Resequencing Study Confirms That Host Defense and Cell Senescence Gene Variants Contribute to the Risk of Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **200**, 199-208 (2019).
21. Baumgartner, K.B. *et al.* Cigarette smoking: A risk factor for idiopathic pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine* **155**, 242-248 (1997).
22. Manna, E.D.F., Pocetta, G., Folletti, I., Paolocci, G. & Rodrigues, C. Environmental and occupational exposures and the risk of idiopathic pulmonary fibrosis: A systematic review. *Giornale Italiano Di Medicina Del Lavoro Ed Ergonomia* **43**, 21-27 (2021).
23. Caminati, A. *et al.* The natural history of idiopathic pulmonary fibrosis in a large European population: the role of age, sex and comorbidities. *Internal and Emergency Medicine*, 10.
24. Garcia-Sancho, C. *et al.* Familial pulmonary fibrosis is the strongest risk factor for idiopathic pulmonary fibrosis. *Respiratory Medicine* **105**, 1902-1907 (2011).
25. Genetics Generation - Nucleotides and Bases. Vol. 2021.
26. Baynes, J. & Dominiczak, M.H. *Medical Biochemistry*, (Mosby, 1999).
27. Handbook of Pharmacogenomics and Stratified Medicine. in *Handbook of Pharmacogenomics and Stratified Medicine* (ed. Padmanabhan, S.) i (Academic Press, San Diego, 2014).
28. Bush, W.S. & Moore, J.H. Chapter 11: Genome-Wide Association Studies. *Plos Computational Biology* **8**, 11 (2012).
29. Verlouw, J.A.M. *et al.* A comparison of genotyping arrays. *European Journal of Human Genetics* **29**, 1611-1624 (2021).
30. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499-511 (2010).
31. Kropski, J.A., Blackwell, T.S. & Loyd, J.E. The genetic basis of idiopathic pulmonary fibrosis. *European Respiratory Journal* **45**, 1717-1727 (2015).
32. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics* **9**, 255-266 (2008).
33. Dhindsa, R.S. *et al.* Identification of a missense variant in SPDL1 associated with idiopathic pulmonary fibrosis. *Communications Biology* **4**, 8 (2021).
34. Allen, R.J. *et al.* Genome-Wide Association Study of Susceptibility to Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **201**, 564-574 (2020).
35. Seibold, M.A. *et al.* A Common MUC5B Promoter Polymorphism and Pulmonary Fibrosis. *New England Journal of Medicine* **364**, 1503-1512 (2011).
36. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

37. Allen, R.J. *et al.* Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *Lancet Respiratory Medicine* **5**, 869-880 (2017).
38. Fingerlin, T.E. *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nature Genetics* **45**, 613-+ (2013).
39. Noth, I. *et al.* Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *Lancet Respiratory Medicine* **1**, 309-317 (2013).
40. Leavy, O.C. *et al.* Proportion of Idiopathic Pulmonary Fibrosis Risk Explained by Known Common Genetic Loci in European Populations. *American Journal of Respiratory and Critical Care Medicine* **203**, 775-778 (2021).
41. Stock, C.J. *et al.* Mucin 5B promoter polymorphism is associated with idiopathic pulmonary fibrosis but not with development of lung fibrosis in systemic sclerosis or sarcoidosis. *Thorax* **68**, 436-441 (2013).
42. Borie, R. *et al.* The MUC5B Variant Is Associated with Idiopathic Pulmonary Fibrosis but Not with Systemic Sclerosis Interstitial Lung Disease in the European Caucasian Population. *Plos One* **8**(2013).
43. Halu, A. *et al.* Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. *Human Molecular Genetics* **28**, 2352-2364 (2019).
44. Peljto, A.L. *et al.* Association Between the MUC5B Promoter Polymorphism and Survival in Patients With Idiopathic Pulmonary Fibrosis. *Jama-Journal of the American Medical Association* **309**, 2232-2239 (2013).
45. Dudbridge, F. *et al.* Adjustment for index event bias in genome-wide association studies of subsequent events. *Nature Communications* **10**(2019).
46. Oertelt-Prigione, S. & Mariman, E. The impact of sex differences on genomic research. *International Journal of Biochemistry & Cell Biology* **124**(2020).
47. Zaman, T. *et al.* Differences in Clinical Characteristics and Outcomes Between Men and Women With Idiopathic Pulmonary Fibrosis A Multicenter Retrospective Cohort Study. *Chest* **158**, 245-251 (2020).
48. Han, M.K. *et al.* Sex differences in physiological progression of idiopathic pulmonary fibrosis. *European Respiratory Journal* **31**, 1183-1188 (2008).
49. Gribbin, J. *et al.* Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the UK. *Thorax* **61**, 980-985 (2006).
50. Mannino, D.M., Etzel, R.A. & Parrish, R.G. Pulmonary fibrosis deaths in the United States, 1979-1991 - An analysis of multiple-cause mortality data. *American Journal of Respiratory and Critical Care Medicine* **153**, 1548-1552 (1996).
51. Turnerwarwick, M., Burrows, B. & Johnson, A. CRYPTOGENIC FIBROSING ALVEOLITIS - CLINICAL-FEATURES AND THEIR INFLUENCE ON SURVIVAL. *Thorax* **35**, 171-180 (1980).
52. Peters, S.A.E., Huxley, R.R. & Woodward, M. Do smoking habits differ between women and men in contemporary Western populations? Evidence from half a million people in the UK Biobank study. *Bmj Open* **4**(2014).
53. Meyer, J.D., Holt, D.L., Cherry, N.M. & McDonald, J.C. SWORD '98: surveillance of work-related and occupational respiratory disease in the UK. *Occupational Medicine-Oxford* **49**, 485-489 (1999).

54. Burdorf, A., Jarvholm, B. & Siesling, S. Asbestos exposure and differences in occurrence of peritoneal mesothelioma between men and women across countries. *Occupational and Environmental Medicine* **64**, 839-842 (2007).
55. Assayag, D., Morisset, J., Johannson, K.A., Wells, A.U. & Walsh, S.L.F. Patient gender bias on the diagnosis of idiopathic pulmonary fibrosis. *Thorax* **75**, 407-412 (2020).
56. Quanjer, P.H. *et al.* Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *European Respiratory Journal* **40**, 1324-1343 (2012).
57. Townsend, E.A., Miller, V.M. & Prakash, Y.S. Sex Differences and Sex Steroids in Lung Health and Disease. *Endocrine Reviews* **33**, 1-47 (2012).
58. Pandit, P., Perez, R.L. & Roman, J. Sex-Based Differences in Interstitial Lung Disease. *American Journal of the Medical Sciences* **360**, 467-473 (2020).
59. Tofovic, S.P., Zhang, X.C., Jackson, E.K., Zhu, H. & Petrusevska, G. 2-methoxyestradiol attenuates bleomycin-induced pulmonary hypertension and fibrosis in estrogen-deficient rats. *Vascular Pharmacology* **51**, 190-197 (2009).
60. Redente, E.F. *et al.* Age and sex dimorphisms contribute to the severity of bleomycin-induced lung injury and fibrosis. *American Journal of Physiology-Lung Cellular and Molecular Physiology* **301**, L510-L518 (2011).
61. Ober, C., Loisel, D.A. & Gilad, Y. Sex-specific genetic architecture of human disease. *Nature Reviews Genetics* **9**, 911-922 (2008).
62. Bernabeu, E. *et al.* Sex differences in genetic architecture in the UK Biobank. *Nature Genetics* **53**, 1283-+ (2021).
63. Fawcett, K.A. *et al.* Variants associated with HHIP expression have sex-differential effects on lung function. *Wellcome open research* **5**, 111-111 (2020).
64. Gilks, W.P., Abbott, J.K. & Morrow, E.H. Sex differences in disease genetics: evidence, evolution, and detection. *Trends in Genetics* **30**, 453-463 (2014).
65. Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244-+ (2017).
66. Dobyns, W.B. *et al.* Inheritance of most X-linked traits is not dominant or recessive, just X-linked. *American Journal of Medical Genetics Part A* **129A**, 136-143 (2004).
67. King, T.E. *et al.* Idiopathic pulmonary fibrosis - Relationship between histopathologic features and mortality. *American Journal of Respiratory and Critical Care Medicine* **164**, 1025-1032 (2001).
68. Wells, A.U. *et al.* Idiopathic pulmonary fibrosis - A composite physiologic index derived from disease extent observed by computed tomography. *American Journal of Respiratory and Critical Care Medicine* **167**, 962-969 (2003).
69. du Bois, R.M. *et al.* Ascertainment of Individual Risk of Mortality for Patients with Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **184**, 459-466 (2011).
70. Ley, B. *et al.* A Multidimensional Index and Staging System for Idiopathic Pulmonary Fibrosis. *Annals of Internal Medicine* **156**, 684-U58 (2012).
71. Mura, M. *et al.* Predicting survival in newly diagnosed idiopathic pulmonary fibrosis: a 3-year prospective study. *European Respiratory Journal* **40**, 101-109 (2012).

72. Herazo-Maya, J.D. *et al.* Validation of a 52-gene risk profile for outcome prediction in patients with idiopathic pulmonary fibrosis: an international, multicentre, cohort study. *Lancet Respiratory Medicine* **5**, 857-868 (2017).
73. Huang, Y. *et al.* A functional genomic model for predicting prognosis in idiopathic pulmonary fibrosis. *Bmc Pulmonary Medicine* **15**(2015).
74. Hosein, K.S., Sergiacomi, G., Zompatori, M. & Mura, M. The CALIPER-Revised Version of the Composite Physiologic Index is a Better Predictor of Survival in IPF than the Original Version. *Lung* **198**, 169-172 (2020).
75. Pastre, J. *et al.* Development and Validation of a Clinical Diagnostic Scoring System for the Diagnosis of Idiopathic Pulmonary Fibrosis. *Annals of the American Thoracic Society* **18**, 1803-1810 (2021).
76. Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling. *Plos Genetics* **6**, 9 (2010).
77. Choi, S.W., Mak, T.S.H. & O'Reilly, P.F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* **15**, 2759-2772 (2020).
78. Wray, N.R., Goddard, M.E. & Visscher, P.M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* **17**, 1520-1528 (2007).
79. Dudbridge, F. Polygenic Epidemiology. *Genetic Epidemiology* **40**, 268-272 (2016).
80. Euesden, J., Lewis, C.M. & O'Reilly, P.F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466-1468 (2015).
81. Trevethan, R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health* **5**(2017).
82. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *Plos Genetics* **9**, 17 (2013).
83. Delong, E.R., Delong, D.M. & Clarkepearson, D.I. COMPARING THE AREAS UNDER 2 OR MORE CORRELATED RECEIVER OPERATING CHARACTERISTIC CURVES - A NONPARAMETRIC APPROACH. *Biometrics* **44**, 837-845 (1988).
84. Lee, S.H., Goddard, M.E., Wray, N.R. & Visscher, P.M. A Better Coefficient of Determination for Genetic Profile Analysis. *Genetic Epidemiology* **36**, 214-224 (2012).
85. Wand, H. *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**, 211-219 (2021).
86. Lambert, S.A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics* **53**, 420-425 (2021).
87. Zeggini, E., Gloyn, A.L., Barton, A.C. & Wain, L.V. Translational genomics and precision medicine: Moving from the lab to the clinic. *Science* **365**, 1409-1413 (2019).
88. Lewis, C.M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine* **12**, 11 (2020).
89. Konuma, T. & Okada, Y. Statistical genetics and polygenic risk score for precision medicine. *Inflammation and Regeneration* **41**(2021).
90. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications* **10**(2019).
91. Allen, R.J. University of Leicester (2018).
92. Team, R. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna Austria, 2018).

93. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *Plos Medicine* **12**(2015).
94. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* **48**, 1284-1287 (2016).
95. Robin, X. *et al.* pROC: an open-source package for R and S plus to analyze and compare ROC curves. *Bmc Bioinformatics* **12**(2011).
96. Mercaldo, N.D., Lau, K.F. & Zhou, X.H. Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in Medicine* **26**, 2170-2183 (2007).
97. Wakeley, J. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *Trends in Ecology & Evolution* **11**, 158-163 (1996).
98. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 16 (2015).
99. LDproxy Tool. Vol. 2021 Interactively explore proxy and putatively functional variants for a query variant.
100. Machiela, M.J. & Chanock, S.J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555-3557 (2015).
101. Werner, M. *et al.* Asthma is associated with single-nucleotide polymorphisms in ADAM33. *Clinical and Experimental Allergy* **34**, 26-31 (2004).
102. Arthur, L.M. *et al.* Structural and functional analysis of Mre11-3. *Nucleic Acids Research* **32**, 1886-1893 (2004).
103. Hambly, N., Shimbori, C. & Kolb, M. Molecular classification of idiopathic pulmonary fibrosis: Personalized medicine, genetics and biomarkers. *Respirology* **20**, 1010-1022 (2015).
104. Uh, S.T. *et al.* ADAM33 Gene Polymorphisms are Associated with the Risk of Idiopathic Pulmonary Fibrosis. *Lung* **192**, 525-532 (2014).
105. Marees, A.T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research* **27**, 10 (2018).
106. Han, J., Jiang, W., Ye, Y. & Zhao, H. Identifying sex-specific genetic effects across 733 traits in UK Biobank [PREPRINT]. *Nature Portfolio* (2021).
107. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291-+ (2015).
108. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236-+ (2015).
109. Thaler, R.H. Anomalies: The Winner's Curse. *The Journal of Economic Perspectives* **2**, 191-202 (1988).
110. Bowden, J. & Dudbridge, F. Unbiased Estimation of Odds Ratios: Combining Genomewide Association Scans with Replication Studies. *Genetic Epidemiology* **33**, 406-418 (2009).
111. Zollner, S. & Pritchard, J.K. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *American Journal of Human Genetics* **80**, 605-615 (2007).

112. Shi, J.X. *et al.* Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data. *Plos Genetics* **12**, 24 (2016).
113. Goring, H.H.H., Terwilliger, J.D. & Blangero, J. Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics* **69**, 1357-1369 (2001).
114. Vilhjalmsson, B.J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics* **97**, 576-592 (2015).
115. Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* **58**, 267-288 (1996).
116. Prive, F., Aschard, H. & Blum, M.G.B. Efficient Implementation of Penalized Regression for Genetic Risk Prediction. *Genetics* **212**, 65-74 (2019).
117. Pattee, J. & Pan, W. Penalized regression and model selection methods for polygenic scores on summary statistics. *Plos Computational Biology* **16**, 27 (2020).
118. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X.Y. & Sham, P.C. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* **41**, 469-480 (2017).
119. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A. & Smoller, J.W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications* **10**(2019).
120. Newcombe, P.J., Nelson, C.P., Samani, N.J. & Dudbridge, F. A flexible and parallelizable approach to genome-wide polygenic risk scores. *Genetic Epidemiology* **43**, 730-741 (2019).
121. Dudbridge, F. & Newcombe, P.J. Accuracy of Gene Scores when Pruning Markers by Linkage Disequilibrium. *Human Heredity* **80**, 178-186 (2015).
122. Kulm, S., Mezey, J. & Elemento, O. Benchmarking the Accuracy of Polygenic Risk Scores and their Generative Methods. *medRxiv*, 2020.04.06.20055574 (2020).
123. Sun, X. & Xu, W.C. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *Ieee Signal Processing Letters* **21**, 1389-1393 (2014).

APPENDICES

APPENDIX A

Properties of U statistic and comparison of 2 ROC curves. Full details including the determination of the variance and covariance from work by de Long et al.[83] and Sun et al.[123].

Empirical AUC

Consider a sample with *m* cases and *n* controls, with $X_i, \dots, X_m$the values of score for cases, and $Y_j, \dots, Y_n$ the values of score for controls. The empirical sensitivity and specificity at each score (cut-off point) are given by:

$$sensitivity(score) = \frac{1}{m}\sum_{i=1}^{m} 1(X_i \geq score) \tag{A.1}$$

$$specificity(score) = \frac{1}{n}\sum_{j=1}^{n} 1(Y_j < score) \tag{A.2}$$

The empirical ROC curve is produced by calculating the sensitivity and specificity at all calculated values of score and plotting sensitivity(score) against [1 - specificity(score)], with the AUC given below.

$$AUC = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \psi(X_i, Y_j), \qquad \psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases} \tag{A.3}$$

The area under the ROC curve is equal to the Man-Whitney U-statistic. This is an unbiased estimate of the probability that a random control will have a score less than or equal to a random case.

Variance and Covariance of 2 Correlated ROC curves

Let $\hat{\theta} = (\hat{\theta}^1, \dots, \hat{\theta}^k)$ be a vector representing the AUCs of the correlated ROC curves. The cases (X_m) and controls (Y_n) contribute to 2 structural components, V10 and V01, per ROC curve. For the r^{th} ROC curve.V10 and V01 are defined as:

$$V_{10}^r(X_i) = \frac{1}{n}\sum_{j=1}^{n} \psi\left(X_i{}^r,\ Y_j{}^r\right), \qquad (\ i = 1, \dots, m) \tag{A.4}$$

$$V_{01}^r(Y_i) = \frac{1}{m}\sum_{i=1}^{m} \psi\left(X_i{}^r,\ Y_j{}^r\right), \qquad (\ j = 1, \dots, n) \tag{A.5}$$

Define the $k \times k$ (here 2 x 2) matrices S_{10} and S_{01}.

Where s refers to the s^{th} ROC curve and so that the $(r, s)^{th}$ elements are

$$S_{10}^{r,s} = \frac{1}{m-1}\sum_{i=1}^{m} \left[V_{10}^r(X_i) - \ \hat{\theta}^r\ \right]\left[V_{10}^s(X_i) - \ \hat{\theta}^s\ \right] \tag{A.6}$$

And

$$S_{01}^{r,s} = \frac{1}{n-1}\sum_{j=1}^{n} \left[V_{01}^r(Y_i) - \ \hat{\theta}^r\ \right]\left[V_{01}^s(Y_i) - \ \hat{\theta}^s\ \right] \tag{A.7}$$

For 2 correlated ROC curves A and B, the covariance matrix for $\hat{\theta}$, (**S**), is:

$$\boldsymbol{S} = \begin{bmatrix} Var[\hat{\theta}^A] & Cov\left[\hat{\theta}^A, \hat{\theta}^B\right] \\ Cov\left[\hat{\theta}^B, \hat{\theta}^A\right] & Var[\hat{\theta}^B] \end{bmatrix}$$

And is calculated by:

$$S = \frac{1}{m} S_{10} + \frac{1}{n} S_{01} \tag{A.8}$$

Statistical test of 2 ROC Curves

For 2 correlated ROC curves, such as 2 different thresholds tested in the male subjects, a z-score can be calculated as shown below, using the AUC of each ROC curve ($\hat{\theta}^A$ and $\hat{\theta}^B$) along with their variances and the covariance between them:

$$z = \frac{\hat{\theta}^A - \hat{\theta}^B}{\sqrt{Var\left[\hat{\theta}^A - \hat{\theta}^B\right]}} = \frac{\hat{\theta}^A - \hat{\theta}^B}{\sqrt{Var\left[\hat{\theta}^A\right] + Var\left[\hat{\theta}^B\right] - 2Cov\left[\hat{\theta}^A, \hat{\theta}^B\right]}} \tag{A.9}$$

Under the null hypothesis of no difference between the ROC curves (H_0=0) the z-distribution is the standard normal distribution $z \sim N(0,1)$ and the two tailed p-value can be obtained.

Uncorrelated ROC curves using an unpaired t-test with unequal sample size and unequal variance. With variances of each ROC curve calculated as above. Vector $\hat{\theta}$ here contains 1 element and $r = s = 1$, which reduces the covariance estimator to a variance estimator in (A.*8*). This is repeated to obtain the variance for each independent curve.

APPENDIX B

GRS code

```
#-----------------------------------------------------------------#
# Read in and process files
#-----------------------------------------------------------------#
file1 <- read.csv("./data/meta_beta_15_SNPs_v2.csv")
file2 <- read.csv("./data/UUS_data_15_SNP_v2.csv")
file3 <- read.csv("./data/UUS_data_15_SNP_info.csv")

#order by chromosome and position
file3_s<- file3[ order(file3$chromosome, file3$position), ]

#-----------------------------------------------------------------#
# Create frames for PRS
#-----------------------------------------------------------------#

#---------------#
# Merge file 1 & 3 - keeping those in both files - 15 SNPs
#---------------#
SNPs_all <- merge(file1, file3_s,
                  by = c("chromosome", "position", "rsid"))

# add OR for beta - just for information
SNPs_all$OR <- exp(SNPs_all$beta)

#-----------------------------------------------------------------#
# Calculation of PRS
#-----------------------------------------------------------------#
#create rsid + effect SNP combo
SNPs_all$rsid_allele <- paste(SNPs_all$rsid,SNPs_all$effect_allele,
                              sep="_", collapse=NULL)

#---------------#
# Set up vectors for calculation of PRS
#---------------#
#keep rsid and beta
SNP_effect <- rbind(SNPs_all$rsid_allele, SNPs_all$beta)
SNP_effect2 <- SNPs_all$beta

#order columns in UUS data to correspond to effects vector
file2_reorder <- file2[c(SNP_effect[1,1:15])]

#---------------#
#create score - multiply coefficient by number of alleles
#---------------#
PRS_score <- mapply(`*`, file2_reorder, SNP_effect2)
PRS_score <- as.data.frame(PRS_score)
```

```r
#sum totals per individual
PRS_score$score <- (rowSums(PRS_score))/14
score <- PRS_score$score

#---------------#
# add back to UUS data - covariates
#---------------#

PRS <- cbind(file2, score)
save(PRS, file="./data/UUS_15_SNP_score.RData")

# end of program
```

APPENDIX C

PRSice example call: fastscore

```
R --file=/scratch/gen3/rja34/PRSice_v1.25.R -q --args \
plink ./plink \
base /scratch/gen3/afg7/Larger_dataset/GWAS_base.txt \
target /scratch/gen3/afg7/Larger_dataset/uus_rsq0.5_qc_sex_pheno_common \
pheno.file /scratch/gen3/afg7/Larger_dataset/pheno.txt \
covary T \
user.covariate.file /scratch/gen3/afg7/Larger_dataset/IPF.covariate \
covariates sex,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10 \
binary.target T \
stat OR \
clump.snps T \
fastscore T \
report.best.score.only F \
report.individual.scores T \
calculate.abc.r2 T \
n.ca.base 2668 \
n.co.base 8591 \
n.ca.targ 793 \
n.co.targ 9999 \
prev.base 0.00063 \
prev.targ 0.00063 \
figname PRS_fastscore \
debug.mode T
```

APPENDIX D

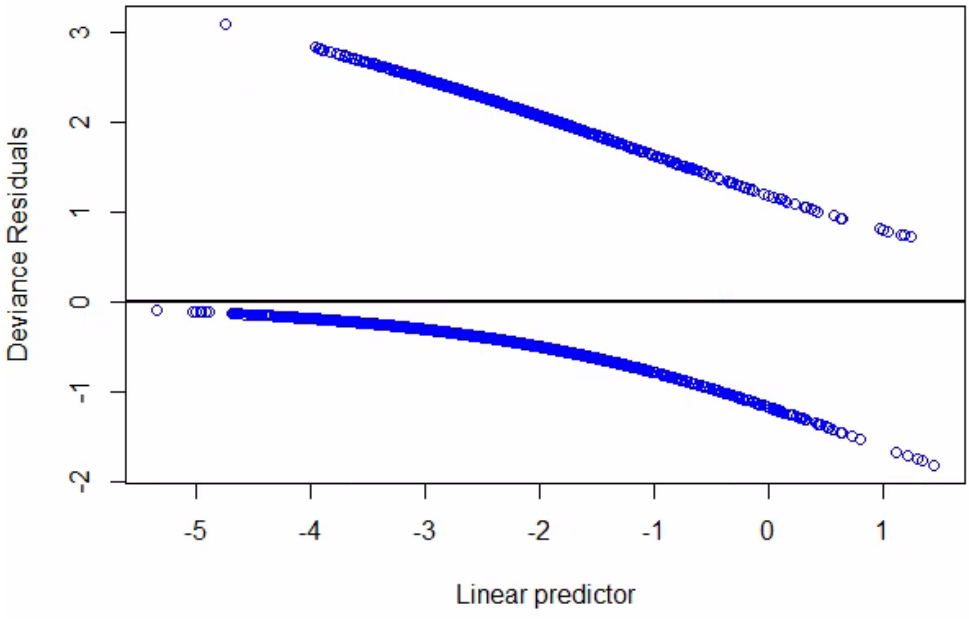


Figure D.1: Deviance Residuals against fitted predictor for PRS at pT 0.1012

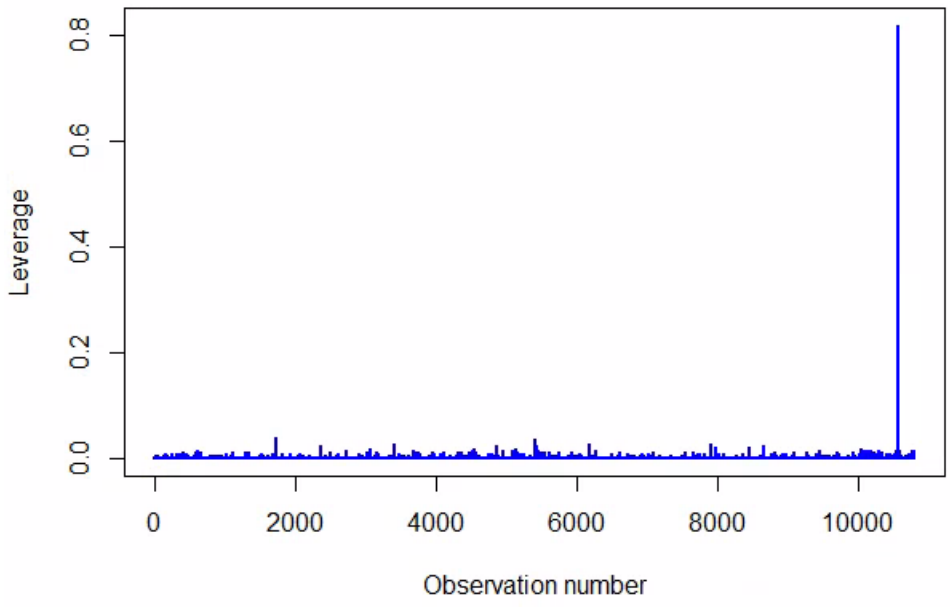


Figure D.2: Leverage of individuals on PRS model fit at pT 0.1012

APPENDIX E

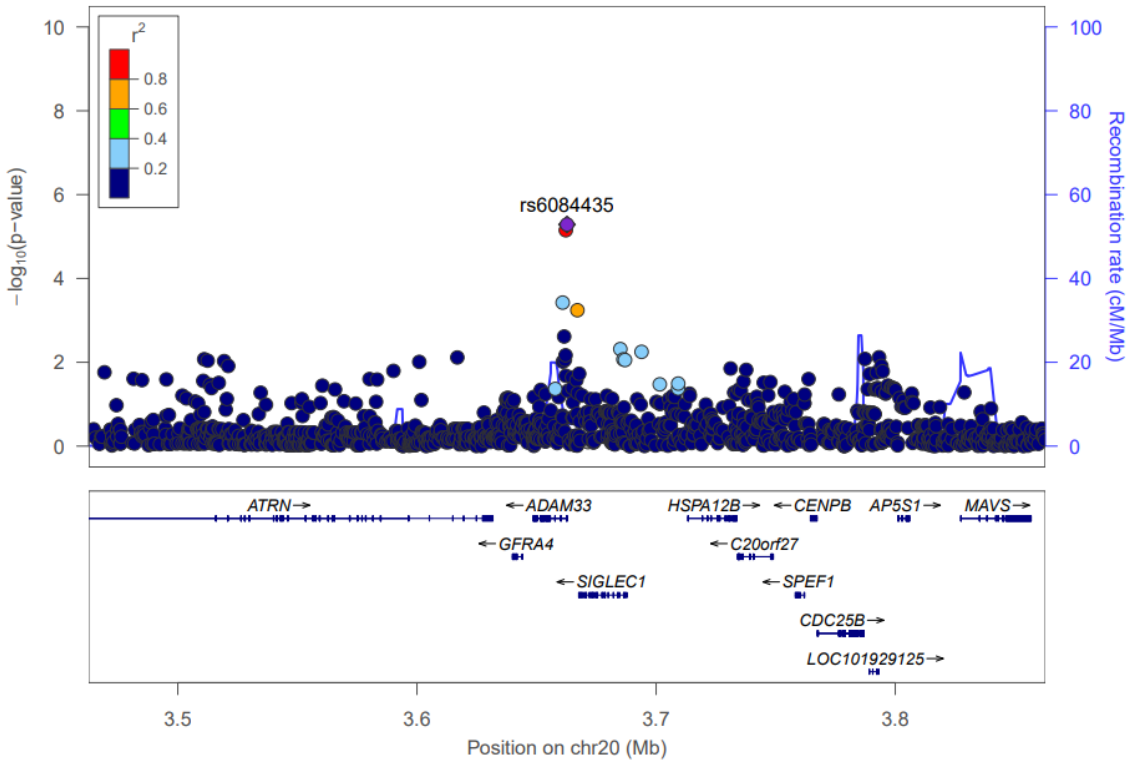


Figure E.3: Region plot of rs6084435. Each point represents a variant with chromosomal position on the x axis and the -log(P value) on the y axis. Variants are colour coded by linkage disequilibrium with rs6084435. Gene locations are shown at the bottom of the plot