

Integrative analysis of common and rare pathogenic variants for a more comprehensive genetic risk assessment

Doctoral thesis

to obtain a doctorate (PhD)

from the Faculty of Medicine

of the University of Bonn

Emadeldin Hassanin

from Cairo, Egypt

2024

Written with authorization of
the Faculty of Medicine of the University of Bonn

First reviewer: Prof. Dr. Peter Krawitz

Second reviewer: Dr. Patrick May

Day or oral examination: 03.09.2024

From the Institute for Genomic Statistics and Bioinformatics

Director: Prof. Dr. Peter Krawitz

Table of Contents

List of abbreviations	4
1. Abstract	5
2. Introduction and aims	6
2.1 Background	6
2.2 Common variants and rare variants	8
2.3 Combining common and rare variants along with family history for complex diseases risk prediction	9
2.4 Generalizability of PRS across diverse ethnicities.....	9
2.5 Aims of the thesis	10
3. Publications	12
3.1 Publication 1: Breast and prostate cancer risk: The interplay of polygenic risk, rare pathogenic germline variants, and family history	13
3.2 Publication 2: Clinically relevant combined effect of polygenic background, rare pathogenic germline variants, and family history on colorectal cancer incidence.....	24
3.3 Publication 3: Assessing the performance of European-derived cardiometabolic polygenic risk scores in South-Asians and their interplay with family history	37
4. Discussion and Conclusion	49
4.1 Impact of polygenic risk scores	49
4.2 Generalizability of European polygenic risk scores to South Asians	51
4.3 Clinical utility and challenges of PRS	51
4.4 Limitations	53
4.5 Future work	53
4.6 Conclusion	54
5. References	55
List of Publications	61
6. Acknowledgements	62

List of abbreviations

AF - Allele Frequency

ATM - Ataxia Telangiectasia Mutated

BOADICEA - Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm

BRCA1 - Breast Cancer Gene 1

BRCA2 - Breast Cancer Gene 2

CAD - Coronary Artery Disease

CHEK2 - Checkpoint Kinase 2

Cox - Cox Proportional Hazards

gnomAD - Genome Aggregation Database

GWAS - Genome-Wide Association Studies

HOXB13 - Homeobox B13

ICD-9 - International Classification of Diseases, 9th Edition

ICD-10 - International Classification of Diseases, 10th Edition

OR - Odds Ratio

PALB2 - Partner and Localizer of *BRCA2*

PCs - Principal Components

PRS - Polygenic Risk Scores

SD - Standard Deviation

SNPs - Single Nucleotide Polymorphisms

T2D - Type 2 Diabetes

1. Abstract

Throughout this thesis, we examine the complex interplay between genetic factors, namely polygenic risk scores (PRS), rare pathogenic variants, and the impact of family history on the risk of complex diseases (e.g., breast and prostate cancer). Using their combined effect on cancer prevalence and lifetime incidence, we aim to demonstrate how they can be used to personalize cancer risk assessment.

For the analyses conducted in this thesis, we have leveraged the large data of UK Biobank. At the time of these analyses, there were 200,643 samples available with both whole exome sequencing and genotyping data. This comprehensive dataset allowed us to classify individuals based on the carrier status of rare pathogenic variants in cancer susceptibility genes, if they have a high or low PRS (defined by 90th and 10th percentile thresholds), and whether they have a family history of cancer. Cox proportional hazards models were used to compute lifetime cumulative incidence of cancer, and multivariate logistic regression was used to compare odds ratios (ORs) across these groups.

Based on genetic profiles of the individuals, the incidences of breast and prostate cancers have shown a distinct variation. For instance, compared to Individuals with lower PRS and absence of rare pathogenic variants, those with rare pathogenic variants and higher PRS exhibit a significantly higher cumulative incidence of cancer by age 70. Further, a family history of respective cancer increases the risk regardless of PRS.

The findings of this thesis highlight the potential use of PRS in risk stratification approaches not only in the general population but also among individuals who carry rare pathogenic mutations. Breast and prostate cancer risks were shown to be influenced both independently and cumulatively by rare pathogenic variants, polygenic background, and family history. The thesis also highlights the urgent need for generalizability of PRS models across diverse ethnic backgrounds to enable more tailored and precise strategies for disease and cancer prevention.

2. Introduction and aims

2.1 Background

Different genetic inheritance models were proposed to explain the contribution of genetics to human traits such as disease status and normal phenotypic variations. Such models can be categorized into two main categories: rare and complex (Rahim et al. 2008). Rare and high heritable traits are usually determined by the presence or absence of a single gene mutation directly causing the observed trait. On the other hand, complex traits which usually account for the majority of phenotypic variations, are driven by a large contribution of polygenic effect where many common variants work together to develop the observed trait (Price et al. 2015).

In genetic studies, the differentiation of common and rare variants is crucial as it is influenced by both technical factors and biological factors. Historically, because of financial limitations in sequencing technology, there was a prioritization of identifying common variants with a frequency of more than 1% in the population (Uffelmann et al. 2021). This focus on common variants was driven by the financial constraints of sequencing compared to genotyping arrays (Bhérier et al. 2024). However, it has several setbacks as the effect sizes for majority of the polymorphisms identified in genome-wide association studies (GWAS) are small and typically they are not the disease-causing variants but rather are in linkage with a disease-causing variant. At the same time, rare variant analysis was limited to moderate to high risk genes that are usually identified through linkage analysis in affected families (Povysil et al. 2019). Taking aside the financial barriers, modelling the risk associated with common variants with small effect sizes, needs summation and integration due to limited cohort's size. This is indirectly linked to frequency. The reason of why high-risk variants are rare can be simply explained by evolutionary selection, where selective pressures tend to push these variants towards lower frequencies (Lee et al. 2014). Therefore, understanding and distinguishing between rare and common variants is crucial for a comprehensive assessment of genetic risk factors (Kachuri et al. 2024).

However, many explanations have been proposed a more complex model regarding what is called "missing heritability" (Eichler et al. 2010). With GWAS primarily considering common variants, latest studies showed that investigating low-frequency

and rare variants could explain further trait or disease risk variability, suggesting a much more sophisticated genetic basis, involving multiple genetic variations (Gibson 2012). In particular, multiple studies showed that both common genetic variants with small additive effects and rare variants with larger effects contribute to the genetic risk of common psychiatric disorders, such as schizophrenia, bipolar disorder, and autism (Weiner et al. 2017; Toma et al. 2018). In the same context, research studies on cancers with strong familial inheritability patterns have shown that the overall genetic risk can be referred to the combined effects of both common variants and rare pathogenic mutations (Lee et al. 2019). Previously, it had been thought that in rare monogenic variants in disorders like neurodevelopmental disorders show complete penetrance, However, recent studies revealed that polygenic effects can also contribute to the phenotypic variance, disease onset age and symptom severity (Niemi et al. 2018; Kurki et al. 2019). In other words, the genetic architecture of complex diseases is more complex than previously thought, and both common and rare genetic variants play an important role in their etiology (Dornbos et al. 2022; Kessler et al. 2022; Fiziev et al. 2023; Ghouse et al. 2024).

In particular, cancer shows a remarkable global health challenge, with far-reaching consequences for both individual patients and society as a whole. Breast, prostate, and colorectal cancers are among the most widespread types of cancer diagnoses (Siegel et al. 2023; Palshof et al. 2024). The importance of such diseases is emphasized by their widespread occurrence, as well as the complex interplay of genetic susceptibility and environmental factors that drive their development. The high incidence of these cancers determines the urgent need for a deeper understanding of their pathogenesis, as well as the development of more effective interventions. To address this challenge, it is essential to investigate the complex genetic architecture and environmental interactions that contribute to the development and progression of these diseases. This extensive approach is essential to better understand, prevent, and treat these pervasive and life-altering diseases (Gao et al. 2021).

2.2 Common variants and rare variants

Recently, cancer research has contributed toward a significant progress in terms of understanding the genetic determinants of these diseases. Specifically, the two genetic components have been of particular interest: Common variants, and rare pathogenic variants (Mars et al. 2020b; Darst et al. 2021).

2.2.1 Common variants and polygenic risk scores

The Polygenic Risk Score (PRS) has been proposed as a potential tool in genetic epidemiology, especially in common diseases such as cancer. PRS involves the integration of the effect size of multiple genetic variants typically taken from GWAS studies to generate a comprehensive risk score (Roberts et al. 2023; Tamlander et al. 2024). This score may be used to provide a more precise assessment of susceptibility to disease. Compared to traditional single-gene approaches, PRS considers the cumulative effects of several genetic variants, each with a small effect size but collectively contributing to the overall risk of disease development (Khera et al. 2018; Mars et al. 2020a).

To identify and validate these significant genetic variants, GWAS has a crucial role in performing large-scale association research and utilizing advanced genomic techniques (Uffelmann et al. 2021). These variants are then integrated into a PRS algorithm, which specifies every variant with a weighted score indicating its individual contribution to the overall risk. This cumulative score allows for the stratification of individuals into distinct risk categories, enabling personalized risk assessments (Choi and O'Reilly 2019).

2.2.2 Rare variants

Rare pathogenic variants have a large effect and significant impact on developments of various cancers (Susswein et al. 2016). Several genes have been identified to be significantly associated with cancer risk and progression. Taking breast cancer as an example, both *BRCA1* and *BRCA2* genes are characterized as high-risk genes, and have been extensively investigated in the context of hereditary breast cancer research due to their significant impact on disease development (Kuchenbaecker et al. 2017;

Breast Cancer Association Consortium 2022). Individuals carrying rare variants in these genes are at a significantly higher risk of developing breast cancer, hence classifying them as hereditary or familial cases. In addition, other genes like *PALB2*, *CHEK2*, and *ATM* have been identified as intermediate or moderate-risk genes (Easton et al. 2015). The presence of such rare pathogenic variants in breast cancer susceptible genes can substantially contribute towards the risk of developing the disease, classifying them as hereditary or familial cases.

2.3 Combining common and rare variants along with family history for complex diseases risk prediction

The impact of the combined effect of both rare pathogenic and common variants in the form of PRS, and their individual contributions to cancer risk have become a subject of intense investigation (Gao et al. 2021). The aim of this thesis was to investigate the joint contribution of common pathogenic variants to cancer susceptibility, and further to compare this risk quantified by PRS to the risk posed by single rare pathogenic variants in genes associated with high to moderate cancer predisposition. Furthermore, the role of family history in breast cancer risk has also been a matter of study (Mars et al. 2022). A family history of for instance breast cancer is recognized as an important risk factor, suggesting a potential shared genetic predisposition. Providing a better understanding of family history role along with other genetic components is crucial for comprehensive cancer risk assessment (Figure 1).

2.4 Generalizability of PRS across diverse ethnicities

Apart from studying the role of rare pathogenic variants and PRS, generalizing the findings across diverse ethnic groups is still a challenging concern (Graham et al. 2021). Many of the genetic risk models and PRS are developed using data primarily from European populations. This poses several concerns in terms of accuracy and effectiveness when it comes to applying these models on non-European populations where genetic backgrounds can differ substantially (Duncan et al. 2019). Ancestry-specific genetic differences in linkage disequilibrium, risk variants, effect sizes, and allele frequencies can hinder the applicability of PRS derived from one population to another (Martin et al. 2017). Such a limitation has been a subject of concern, as most

large-scale GWAS have primarily focused on individuals of European ancestry, potentially limiting the usefulness of these PRS in non-European populations (Privé et al. 2022).

One such example of underrepresented population in genetic studies is South Asians (SAS), even though they constitute a substantial portion of the global population (Huang et al. 2022). SAS individuals are known to have an increased susceptibility to coronary artery disease (CAD), obesity, and type 2 diabetes (T2D). There are two potential approaches to address this problem: 1) Generate more diverse genetic data sets including individuals from SAS population (Wang et al. 2020), or 2) Develop statistical methods to make use of the PRSs derived from European population (Ge et al. 2019). In the short term, the second option is more feasible and cost-effective. Further, similar to the European population. Further, understanding the utility of PRS in SAS individuals and exploring their interplay with family history in disease risk prediction are critical research areas (Hujuel et al. 2022).

2.5 Aims of the thesis

In this thesis, we investigate two major aspects of genetics research:

1. The complex interplay between rare pathogenic variants, polygenic background, and family history in breast, prostate, and colorectal cancers. Investigating the cumulative and independent contributions of these factors to cancer risk will help us understand their combined impact.
2. Transferability of PRS between different ancestral populations in clinical settings as it remains a significant concern. The purpose of this study is to assess whether European-derived-PRSs can be transferred to different ethnic groups, including South and East Asians, and to evaluate feasibility of implementating the PRS in clinical settings.

Genetic overview of breast cancer

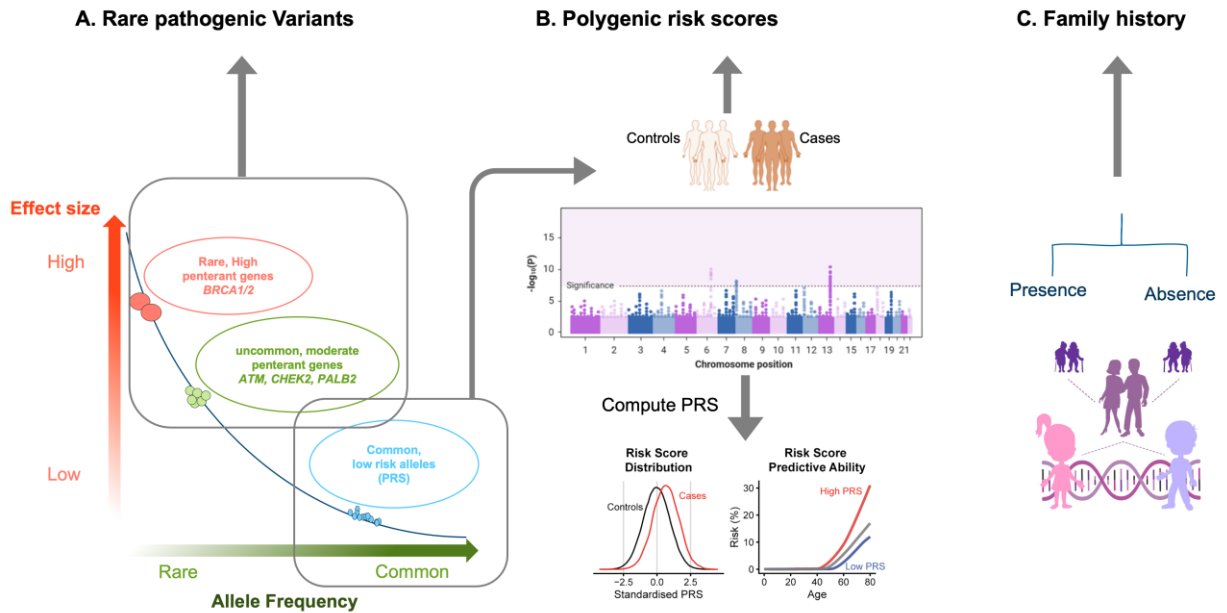


Figure 1 An illustrative example of complex genetic architecture of breast cancer showing the role of various genetic risk factors and family history in developing the disease.

3. Publications

3.1 Publication 1: Breast and prostate cancer risk: The interplay of polygenic risk, rare pathogenic germline variants, and family history

Contribution: Data analysis, interpretation of results, writing and revision of manuscript



ARTICLE

Breast and prostate cancer risk: The interplay of polygenic risk, rare pathogenic germline variants, and family history



Emadeldin Hassanin¹, Patrick May², Rana Aldisi¹, Isabel Spier^{3,4}, Andreas J. Forstner^{3,5,6}, Markus M. Nöthen³, Stefan Aretz^{3,4}, Peter Krawitz¹, Dheeraj Reddy Bobbili², Carlo Maj^{1,*}

¹Institute for Genomic Statistics and Bioinformatics, University of Bonn, Bonn, Germany; ²Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg; ³Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany; ⁴National Center for Hereditary Tumor Syndromes, University Hospital Bonn, Bonn, Germany; ⁵Centre for Human Genetics, Philipps-University Marburg, Marburg, Germany; ⁶Institute of Neuroscience and Medicine (INM-1), Research Center Jülich, Jülich, Germany

ARTICLE INFO

Article history:

Received 6 June 2021

Received in revised form

12 September 2021

Accepted 12 November 2021

Available online 18 November 2021

Keywords:

Breast cancer

Family history

Polygenic risk score

Prostate cancer

Rare pathogenic variants

ABSTRACT

Purpose: We aimed to investigate to what extent polygenic risk scores (PRS), rare pathogenic germline variants (PVs), and family history jointly influence breast cancer and prostate cancer risk.

Methods: A total of 200,643 individuals from the UK Biobank were categorized as follows: (1) heterozygotes or nonheterozygotes for PVs in moderate to high-risk cancer genes, (2) PRS strata, and (3) with or without a family history of cancer. Multivariable logistic regression and Cox proportional hazards models were used to compute the odds ratio across groups and the cumulative incidence through life.

Results: Cumulative incidence by age 70 years among the nonheterozygotes across PRS strata ranged from 9% to 32% and from 9% to 35% for breast cancer and prostate cancer, respectively. Among the PV heterozygotes it ranged from 20% to 48% in moderate-risk genes and from 51% to 74% in high-risk genes for breast cancer, and it ranged from 30% to 59% in prostate cancer risk genes. Family history was always associated with an increased cancer odds ratio.

Conclusion: PRS alone provides a meaningful risk gradient leading to a cancer risk stratification comparable to PVs in moderate risk genes, whereas acts as a risk modifier when considering high-risk genes. Including family history along with PV and PRS further improves cancer risk stratification.

© 2021 The Authors. Published by Elsevier Inc. on behalf of American College of Medical Genetics and Genomics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Dheeraj Reddy Bobbili and Carlo Maj contributed equally.

*Correspondence and requests for materials should be addressed to Carlo Maj, Institute for Genomic Statistics and Bioinformatics, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. E-mail address: cmaj@uni-bonn.de

doi: <https://doi.org/10.1016/j.gim.2021.11.009>

1098-3600/© 2021 The Authors. Published by Elsevier Inc. on behalf of American College of Medical Genetics and Genomics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

Breast cancer and prostate cancer represent 2 of the most common cancers in women and men, respectively. Within the UK Biobank (UKB) cohort, breast cancer is the most prevalent cancer diagnosis in females, and prostate cancer is the most prevalent cancer diagnosis in males (<https://biobank.ctsu.ox.ac.uk/~bbdata/CancerSummaryReport.html>). Along with several other factors, predisposing genetic variants (constitutional/germline variants) play a crucial role in the risk of developing breast cancer and prostate cancer.

Both breast cancer and prostate cancer are characterized by a high heritability, estimated to be around 31% for breast cancer¹ and 58% for prostate cancer.² Within breast cancer cases, approximately 5% to 10% are monogenic forms caused by moderate to high penetrant pathogenic germline variants.³ Similarly, in prostate cancer familial subtypes following a Mendelian inheritance have been identified.⁴ It is noteworthy that in 17% of the patients with family history for prostate cancer, who were referred for genetic testing, a pathogenic germline variant could be identified.⁵ Breast cancer and prostate cancer share some susceptibility genes suggesting a potential shared genetic predisposition between the 2 cancer types.⁶ It has also been observed that family history in first-degree relatives for prostate cancer increases women's risk of developing breast cancer by 14%.⁷ Similarly, having a first-degree relative with breast cancer increases the chance of developing prostate cancer by 18%,⁸ which further underpins the hypothesis of shared genetic risk factors.

Several studies have shown the crucial role of predisposing germline variants in the etiology of breast cancer: rare high-risk variants in *BRCA1* and *BRCA2*⁹; rare intermediate/moderate-risk variants in *PALB2*, *CHEK2*, and *ATM*¹⁰; and various common low risk variants.¹¹ In particular, *BRCA1/2* pathogenic variants are most commonly linked to monogenic breast cancer, usually designated as hereditary breast cancer and ovarian cancer.³

In addition to the risk conferred by rare pathogenic variants in the strongly associated genes, different genome-wide association studies (GWAS) have identified hundreds of single-nucleotide variations associated with breast cancer risk. Although each single-nucleotide variation has a negligible effect size, their cumulative effect calculated as polygenic risk score (PRS) contributes significantly to the cancer risk, and it can improve disease risk stratification in the general population.¹² Although it is well-established that both rare and common constitutive variants are associated with breast cancer, only few studies have explored their combined effect and specifically to what extent the polygenic background acts as a risk modifier of monogenic variants of breast cancer.

For instance, the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm model is a comprehensive breast cancer prediction tool incorporating *BRCA1*, *BRCA2*, *PALB2*, *ATM*, and *CHEK2* variants, along

with other risk factors such as family and medical history, lifestyle, and, recently, also PRS.¹³ In a recent study, the impact of PRS on the penetrance of the breast cancer risk variants was assessed for NM_024675.3:c.1592del (rs180177102) in *PALB2* and NM_007194.3:c.1100del (rs555607708) in *CHEK2* in Finnish population¹⁴ and for *BRCA1/2* cancer-associated variants in a previous release of UKB including a smaller cohort of 49,960 individuals with exome-sequencing data.¹⁵

Similarly, different genes are associated with the etiology of prostate cancer, in particular *BRCA1/2*, *ATM*, *CHEK2*, and *HOXB13*.¹⁶⁻¹⁸ Moreover, several studies have shown that for prostate cancer also the cumulative risk driven by the presence of common variants as summarized by PRS models is strongly associated with the cancer risk.¹⁹ Few studies showed the effect of PRS stratification among heterozygotes for p.G84E in *HOXB13*,²⁰ and heterozygotes for *BRCA1/2* pathogenic variant.²¹ However, those studies focused only on specific variants or genes.

In this work, we compared the prevalence and the lifetime risk of breast cancer and prostate cancer among 200,643 individuals from the UKB. Individuals were categorized into heterozygotes and nonheterozygotes of rare pathogenic or likely pathogenic (P/LP) variants (hereafter defined as PV) in moderate or high susceptibility genes; low, intermediate, and high PRS; and with or without a family history for the respective cancer.

Material and Methods

Data source

This study was performed using genetic and phenotypic data from UKB (application number 52446). UKB is a long-term prospective population-based study, and the volunteers are being recruited mainly from England, Scotland, and Wales; it involves more than 500,000 participants aged between 40 and 69 years at recruitment. An abundant diversity of phenotypic and health-related information is available on each participant; for 487,410 samples, genotyping data are available, and for 200,643 individuals, exome sequencing (ES) data are also available. The data set is accessible for research purposes, and all participants provided documented consent.²²

Study participants

Breast cancer cases were defined on the basis of self-reported code 1002 (in data field 20001), International Classification of Diseases (ICD)-10 code C50.X, or ICD-9 code 174.X in hospitalization records. For prostate cancer, cases with self-reported code 1044 (in data field 20001), ICD 10 code C61 and D075, or ICD-9 code 185 in

hospitalization records were included. The remaining samples with no other cancer diagnosis were considered as controls. Individuals of all ancestries were included in the analysis. Only individuals with both genotyping and ES data were included ($N = 200,643$). On the basis of the available genotype data, we excluded outliers for heterozygosity or genotype missing rates, putative sex chromosome aneuploidy, and discordant reported sex vs genotypic sex. In the analysis, we included only females for breast cancer and only males for prostate cancer. We excluded 1 from each pair of related individuals if the genetic relationship was closer than the second degree, defined as kinship coefficient > 0.0884 as calculated by the UKB (https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/ukbgen_instruct.html).

Variant selection

Annovar²³ was used to annotate the variant call format files per chromosome from the 200,643 ES data. Variant frequencies were retrieved from the Genome Aggregation Database (gnomAD),²⁴ whereas ClinVar²⁵ annotations were considered to interpret the pathogenicity of germline variants.

The following inclusion criteria were applied to select rare PV in the UKB data: (1) only variants in protein-coding regions of the *BRCA1/2*, *CHEK2*, *ATM*, and *PALB2* genes for breast cancer and *BRCA1/2*, *CHEK2*, *ATM*, and *HOXB13* genes for prostate cancer; (2) allele frequency < 0.005 in at least 1 ethnic subpopulation of gnomAD and also allele frequency < 0.005 in gnomAD overall; (3) not annotated as synonymous, nonframeshift deletion, and nonframeshift insertion; and (4) annotated as P/LP on the basis of ClinVar, ie, if the variant is consistently classified as such or, in case of a conflicting interpretation, if at least 3 P/LP annotations were available without any benign/likely benign classification. A similar variant filtering approach has been applied in a recent analysis aimed at identifying disease causing monogenic variants.¹⁵ Individuals carrying any of the identified variants in the moderate to high penetrant genes in heterozygous or homozygous state were classified as PV heterozygotes. We use the term nonheterozygote to refer to individuals who are not heterozygous for a PV variant.

PRS

To generate the PRS, we used a previously validated PRS for breast cancer and prostate cancer containing 313 and 103 variants, respectively.^{21,26} The PRS was calculated from the UKB genotype data using the PLINK 2.0²⁷ scoring function. We applied a previous approach to minimize variance in PRS distributions across genetic ancestries.²⁸ Specifically, we fit linear regression model using the first 4 ancestry principal components (PCs) in the controls ($PC_PRS = PC1 + PC2 + PC3 + PC4$). The

derived model was applied to predict the PC_PRS over the entire data set. The PC adjusted PRS was calculated by subtracting PC_PRS from the raw PRS (ie, the residual PRSs were computed) and used for the subsequent analyses.

Statistical analysis

Individuals were stratified on the basis of the PRS percentile, presence or absence of PV (ie, heterozygous or non-heterozygous), and family history. We considered the corresponding family history of cancer in parents and siblings as reported by participants (UKB Data-fields: 20110, 20107, 20111). We assigned individuals to low ($<10\%$), intermediate (10%-90%), and high ($>90\%$) PRS where the definition of a high PRS (above the 90th percentile) followed a previous study.¹⁸ The rationale to stratify PRS into 3 risk classes was in line with the hypothesis that PRS is associated with a nonlinear decrease of risk for extremely low PRS and nonlinear increase of risk for extremely high PRS as observed in other studies.¹²

Intermediate PRS, nonheterozygote, and an absent family history corresponded to the large majority of individuals (69.9% and 72.1% for breast cancer and prostate cancer, respectively); therefore, this group was used as a reference to assess cancer prevalence in the population (ie, to compute the odds ratios [ORs]). We performed the analysis considering all genes (ie, heterozygotes of variants in any of the susceptibility genes) and also performed gene-specific analysis. For breast cancer, we stratified between PV heterozygotes in genes characterized by moderate/intermediate penetrance (ie, *ATM*, *CHEK2*, *PALB2*, in the following summarized as moderate) and heterozygotes in highly penetrant genes (ie, *BRCA1/2*) to assess the effect of PRS in the 2 risk groups. In contrast, for prostate cancer, we defined only a single group because there is no clear difference in the penetrance of the included genes. For each group, we computed the OR using a logistic regression model adjusted for age at recruitment and the first 4 PCs. We then predicted the cancer ORs across PRS percentiles from a logistic regression model by considering nonheterozygotes without family history with intermediate PRS as reference and conditioning on the mean of covariates (age and the first 4 PCs).

We estimated the lifetime risk by age 70 years resulting from PV status and the PRS. We fit a Cox proportional hazards model using the R package *survival*. We used age as the time scale representing the time-to-event, considering age at diagnosis in cases and age of last assessment in controls. The model included PV heterozygote status, PRS strata (ie, low, intermediate, high), age, and the first 4 ancestry PCs, whereas adjusted survival curves were plotted with the R package *survminer*. For all statistical analyses, we used R 3.6.3.

Results

Stratification of UKB cohort individuals for cancer prevalence, family history, and genetic risk factors

Within the 200,643 UKB individuals with available genotyping and exome data, we identified 6288 breast cancer cases (3838 prevalent cases and 2450 incident cases) with a mean age at diagnosis of 55.6 years. The remaining 85,903 women with no other cancer diagnosis were considered as controls, and the mean age at last visit was 56.8 years (Supplemental Table 1).

For prostate cancer, a total of 4021 cases (1331 prevalent cases and 2690 incident cases) were identified with a mean age at diagnosis of 64.4 years. The remaining 73,053 men with no other cancer diagnosis were considered as controls, and the mean age at last visit was 57.0 years (Supplemental Table 2).

It is noteworthy that both in breast cancer and prostate cancer, there was a significantly higher proportion of individuals with a family history for cancers not only among heterozygotes of PV in the selected cancer susceptibility genes (OR = 2.09 and 1.62, $P < .01$) but also among individuals with high-PRS (OR = 1.38 and 1.37, $P < .01$) (Tables 1 and 2).

Distribution of PV heterozygotes within the UKB cohort

We identified 1622 heterozygotes of 309 PV in the 5 analyzed breast cancer susceptibility genes ie, *BRCA1/2*, *PALB2*, *CHEK2*, and *ATM*.

In addition, 1492 heterozygotes of 259 PV were found in the 5 considered prostate cancer susceptibility genes, ie, *BRCA1/2*, *ATM*, *CHEK2*, and *HOXB13*. The list of the considered variants, annotations, and number of heterozygotes are available in the Supplemental File 2.

Among the study participants, homozygous PVs were not identified either in breast cancer or in prostate cancer.

PRS distribution within the UKB cohort

The breast cancer and prostate cancer PRSs followed a normal distribution (raw and PC-adjusted PRS are shown in Supplemental Figure 1) and were significantly higher in cases than in controls ($P < .01$) (Supplemental Figure 2).

We observed a nonlinear increase of cancer risk for individuals in the extreme right tail of the PRS distribution and a less evident nonlinear decrease in the left tail (Supplemental Figure 3—disease prevalence by PRS percentile for both breast and prostate cancer). This corroborates the hypothesis that PRS can be used to stratify individuals into risk classes according to a liability threshold model²⁹ (ie, low, intermediate, and high risk).

Interplay between PV heterozygosity and PRS

None of the selected PV was included in the PRS, and thus, they represent an independent genetic signal. We observed that the mean and median of PRS was significantly higher in affected heterozygotes than in unaffected heterozygotes (Supplemental Figure 4).

For breast cancer, we performed a separate analysis for the high-risk genes *BRCA1/2* and the moderate-risk genes *PALB2*, *CHEK2*, and *ATM*. We estimated how breast cancer risk is influenced by PRS and the heterozygous status for PV in cancer susceptibility genes by computing the ORs for cancer across groups with respect to nonheterozygotes with intermediate PRS because they represent the major group in the population. Heterozygotes with intermediate PRS represent the heterozygotes population, and therefore, they are designated as heterozygotes for simplicity. The high-risk genes PV heterozygotes had a higher OR than individuals with only a high PRS (5.9 vs 2.0, Figure 1A). Instead, PV heterozygotes in the moderate risk genes had an OR comparable with the OR in case of nonheterozygotes with high PRS (OR = 2.2 vs 2.0), but the number of nonheterozygote women with high PRS was considerably larger than the number of heterozygotes (Figure 1A). Notably, women heterozygous for PV in moderate risk genes (ie, *ATM*, *CHEK2* and *PALB2*) with low PRS had a lower risk than nonheterozygote women with only high PRS (OR 1.2 vs 2.0).

In general, PRS modifies the penetrance of PVs in both moderate- and high-risk genes. Of note, PV heterozygote women with low PRS in case of both high-risk and moderate-risk genes had lower ORs (ie, 2.9 and 1.2, respectively), whereas heterozygote women with high PRS had the largest absolute ORs (OR = 8.6 and 3.3, respectively; Figure 1A).

For prostate cancer, PV heterozygotes with intermediate PRS had OR comparable with that of nonheterozygotes with high-PRS (OR = 2.3 vs 2.2) and even lower in case of low PRS (OR = 1.6). Notably, similar to the number observed in women for breast cancer, the number of nonheterozygote men with high PRS was considerably larger than the number of heterozygotes (Figure 1C). As expected, among PV heterozygotes, men with low PRS had the lowest ORs and the men with high PRS had the highest ORs (1.6 and 6.1, respectively, Figure 1C).

Similarly, analysis of the lifetime cancer risk showed a joint effect of PV and PRS. The cumulative incidence by age 70 years in heterozygotes was the lowest in case of low PRSs and the highest in the case of high PRS. In breast cancer, values ranged from 51% to 74% for high-risk genes and from 20% to 48% for moderate-risk genes (Figure 1B), whereas for prostate cancer the incidence ranged from 30% to 59% (Figure 1D). Notably, for nonheterozygotes the cumulative incidence ranged between 9% and 32% for breast cancer and between 9% and 35% for prostate cancer.

Table 1 Characteristics of the participants categorized by PV heterozygosity status and PRS strata in prostate cancer

	Heterozygote and High PRS	Heterozygote and Intermediate PRS	Heterozygote and Low PRS	Nonheterozygote and High PRS	Nonheterozygote and Intermediate PRS	Nonheterozygote and Low PRS
Participants, <i>n</i>	187	1185	120	7520	60,474	7588
Cases, <i>n</i> (%)	42 (22.46)	118 (9.96)	8 (6.67)	728 (9.68)	2971 (4.91)	154 (2.03)
Controls, <i>n</i>	145 (77.54)	1067 (90.04)	112 (93.33)	6792 (90.32)	57,503 (95.09)	7434 (97.97)
Age, ^a mean (SD)	57.89 (8.92)	57.19 (8.68)	56.31 (8.41)	57.31 (8.73)	57.43 (8.7)	57.31 (8.75)
Family history of prostate cancer, <i>n</i> (%)	33 (17.65)	135 (11.39)	19 (15.83)	798 (10.61)	4880 (8.07)	455 (6)

PRS, polygenic risk score; PV, pathogenic variant.

^aAge at diagnosis for cases and age at last visit for controls.

Inclusion of family history on the cancer risk stratification

A family history of the corresponding cancer was present in 19% and 16% of cases and 10.7% and 7.8% of controls (OR = 2.0 and 2.3, $P < .01$) for breast cancer and prostate cancer, respectively (Supplemental Tables 1 and 2). Considering individuals with no family history and intermediate PRS as reference, we found that both family history and PRS were associated with higher risk (see ORs in Supplemental Figures 5 and 6). The risk was lowest for low PRS and no family history (ORs of 0.45 and 0.42 for breast cancer and prostate cancer, respectively) and the highest in the presence of both family history and high PRS (ORs of 3.5 in breast cancer and 4.6 in prostate cancer).

The full models considering the underlying continuous distribution of PRS by computing the predicted ORs across PRS percentiles in individuals stratified for family history and PV status in moderate-risk and high-risk genes showed that the cancer risk is strongly influenced by PRS in all groups (Figure 2). Considering the nonheterozygotes with no family history and median PRS percentile group as reference, the predicted breast cancer ORs in the lower tail of PRS was 0.36 for nonheterozygotes with no family history, whereas in the upper tail of PRS, for PV heterozygotes with family history, the OR reached 6.6 and 10.3 in

moderate-risk and high-risk genes, respectively. A similar trend was observed for prostate cancer in which the lowest predicted OR of 0.3 was reached for PV nonheterozygotes without family history and OR of 13.1 for heterozygotes with family history and high PRS.

The effect of PRS in single gene heterozygotes

We estimated how PRS influences breast cancer prevalence among PV heterozygote women in each of the analyzed susceptibility genes.

The gene-specific analysis revealed a strong variability in risk conferred by rare PV in different genes. In particular, for breast cancer, the largest effect sizes were attributable to *BRCA1/2*, a comparably lower effect size was present for *PALB2* and *ATM*, and the lowest effect size was observed for *CHEK2* (Supplemental Figure 7). Gene-specific analysis in prostate cancer also showed heterogeneity across gene effect sizes with the largest effect observed for *HOXB13* and the smallest effect observed for *BRCA1* (Supplemental Figure 8). Despite having single genes, both breast and prostate cancers were characterized by different effect sizes, and the PRS modifies the relative risk across all genes.

Similar to the overall analysis, the gene-specific analysis showed that family history, PV, and PRS are associated with increased cancer risk. Despite the genes characterized by

Table 2 Characteristics of the participants by PV heterozygosity status and PRS strata in breast cancer

	Heterozygote and High PRS	Heterozygote and Intermediate PRS	Heterozygote and Low PRS	Nonheterozygote and High PRS	Nonheterozygote and Intermediate PRS	Nonheterozygote and Low PRS
Participants, <i>n</i>	222	1279	121	8997	72,473	9099
Cases, <i>n</i> (%)	56 (25.23)	241 (18.84)	16 (13.22)	1068 (11.87)	4634 (6.39)	273 (3)
Controls, <i>n</i>	166 (74.77)	1038 (81.16)	105 (86.78)	7929 (88.13)	67,839 (93.61)	8826 (97)
Age, ^a mean (SD)	55.51 (8.83)	56.03 (8.7)	54.74 (9.52)	56.52 (8.35)	56.75 (8.4)	57.08 (8.31)
Family history of breast cancer, <i>n</i> (%)	56 (25.23)	260 (20.33)	20 (16.53)	1317 (14.64)	8023 (11.07)	710 (7.8)

PRS, polygenic risk score; PV, pathogenic variant.

^aAge at diagnosis for cases and age at last visit for controls.

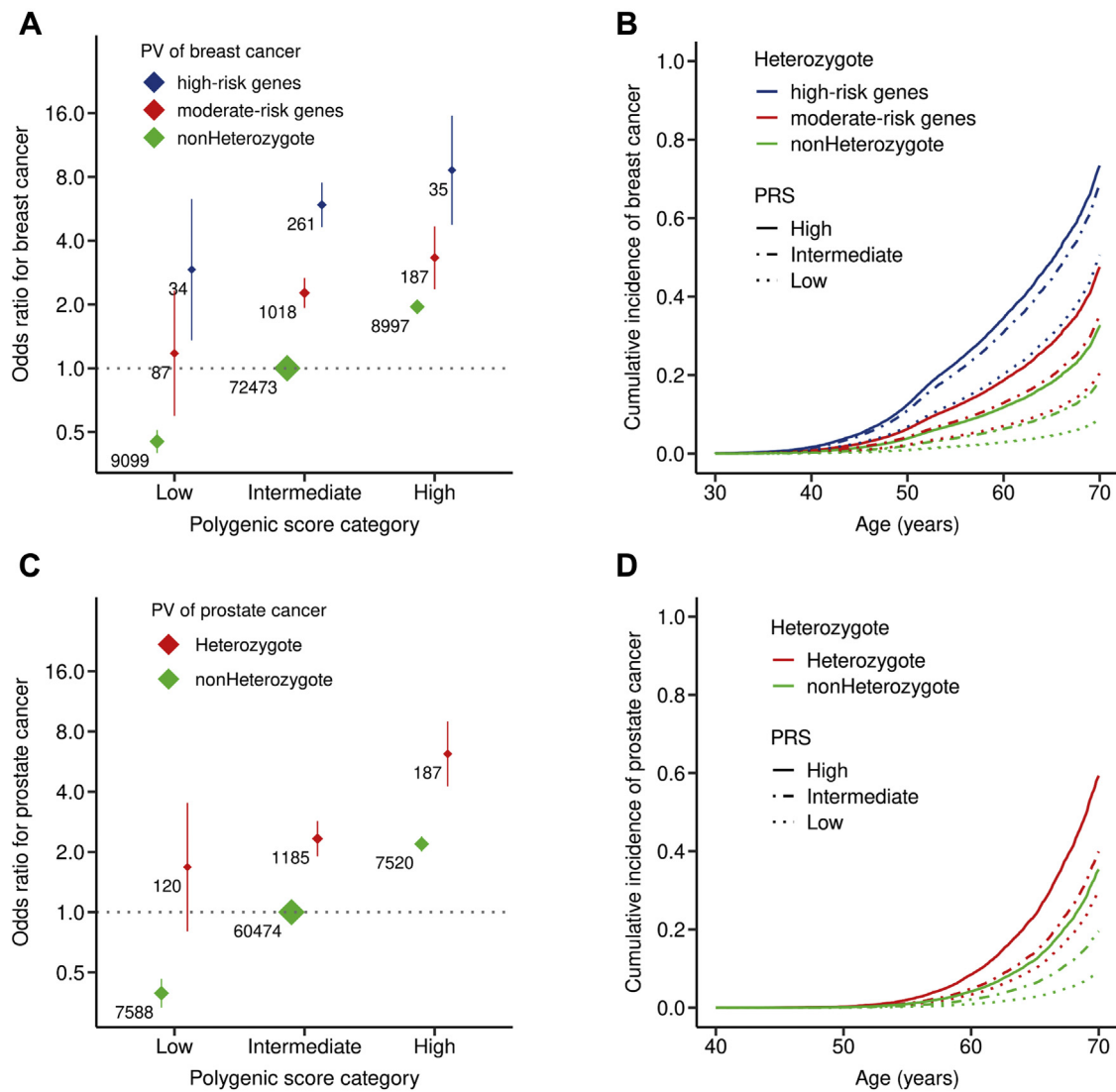


Figure 1 Cancer odds ratio and cumulative incidence among individuals categorized according to the presence of PV heterozygotes and PRS. Heterozygotes and nonheterozygotes were categorized into 3 strata on the basis of their PRS: low (<10 percentile), intermediate (10-90 percentile), or high (>90 percentile) PRS. The odds ratio was calculated from a logistic regression model with age, and the first 4 principal components of ancestry as covariates for breast cancer (A), and prostate cancer (C). The reference group was nonheterozygotes with intermediate PRS. The adjusted odds ratio is indicated by the colored boxes. The numbers next to the odds ratios indicate the sample size of the corresponding group. The 95% CI are indicated by the vertical lines around the boxes. Cumulative incidence was estimated from a Cox proportional hazards model using age, and the first 4 ancestry principal components for breast cancer (B), and prostate cancer (D). PRS, polygenic risk score; PV, pathogenic variant.

different risk levels, family history lead to larger ORs, and this trend was observed across different PRS strata (Figure 3A and B for breast and prostate cancer, respectively).

Discussion

In this study, we analyzed how breast and prostate cancer prevalence and cumulative incidence within the UKB cohort is affected by genetic susceptibility and family history. We considered both the genetic component driven by rare PV in genes associated with hereditary forms of cancer and the polygenic background present in all individuals.

Our results support the hypothesis of cumulative genetic risks caused by both rare PV and the polygenic background. We observed a higher prevalence of cancer in PV heterozygotes with high PRS (ie, individuals with suspected hereditary forms of breast cancer and prostate cancer). This result corroborates the role of the polygenic background as a modifier of the breast cancer and prostate cancer risk among PV heterozygotes unselected for specific clinical criteria (as the UKB cohort), and this is in line with that observed in other studies focused on specific genes or variants.^{14,15} Lifetime risk analysis of breast cancer and prostate cancer indicated that the cumulative disease incidence can be jointly influenced by the presence of PV and the polygenic contribution over the course of life.

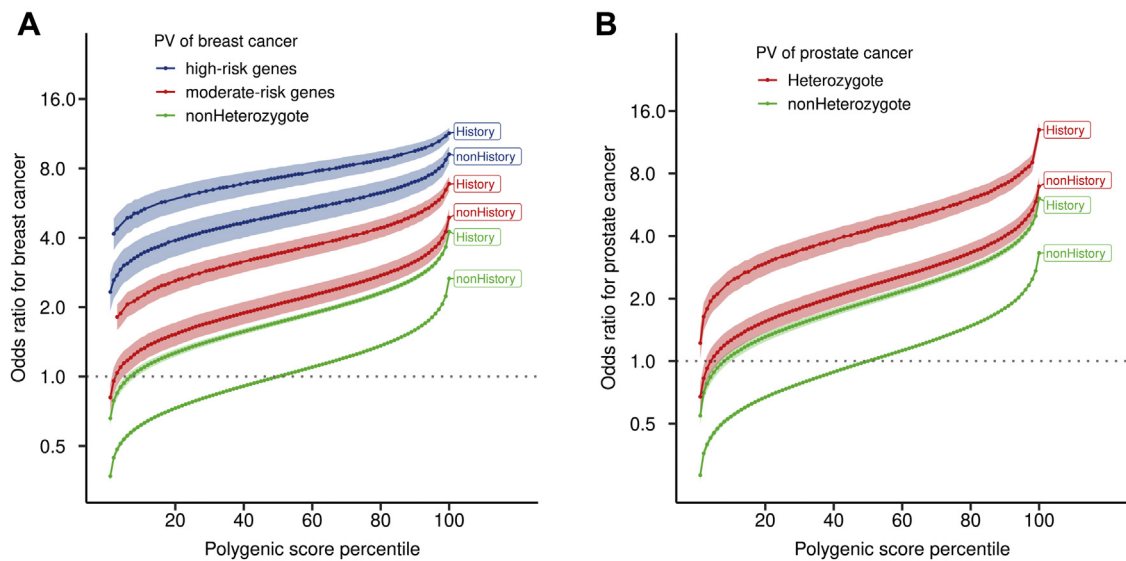


Figure 2 Interplay of PV, family history, and polygenic risk score (PRS). Predicted odds ratios for cancer were estimated from logistic models adjusted for age and first 4 ancestry principal components for breast cancer (A), and for prostate cancer (B). Nonheterozygotes with median PRS and no family history served as the reference group. PRS, polygenic risk score; PV, pathogenic variant.

Single-gene analysis revealed heterogeneous effects across genes, and therefore, the modifier role exerted by PRS should be framed within the absolute risk attributable to individual genes. This is in line with a recent study suggesting that PRS inclusion in risk stratification may prevent excess of surveillance for breast cancer in PV heterozygotes in moderate-risk genes such as *CHEK2* and *ATM*, whereas the cancer risk for PV heterozygotes in high-risk genes such as *BRCA1/2* is clinically relevant irrespective of the PRS.³⁰ Another recent work showed that there is a wide-range of absolute risks for breast cancer and prostate cancer in PV heterozygotes in terms of different genes and across PRS stratification.³¹

Our results showed that the PRS acts as a risk modifier for breast cancer and prostate cancer among both the general population and PV heterozygotes in all the well-known cancer susceptibility risk genes. PRS can define a significant proportion of the general population that is at a risk comparable with PV heterozygotes for moderate-risk genes or even more when considering family history. According to these findings, there should be a potential benefit including PRS in health care prevention policies for both the general population and at-risk individuals carrying PVs because risk-stratified surveillance might improve early disease detection and prevention.^{32,33}

In particular, we observed that women with PVs in moderate-risk genes *ATM*, *CHEK2*, or *PALB2* with a high PRS had a cumulative incidence comparable with women with PV in high-risk genes *BRCA1/2* with a low PRS. On the contrary, women heterozygous for PV in moderate-risk genes with a low PRS had a cumulative incidence comparable to the general population. These results suggest that for women with PV in moderate-risk genes, the addition of PRS can optimize the risk stratification, which is often based

on the life-time risk. Therefore, especially in the presence of PV in moderate-risk genes for breast cancer, intense surveillance programs and potential preventive measures can be better assessed when including the modifier role of PRS.

Moreover, with increasing population-based cohort sizes, PRS can better define a small group of very high-risk nonheterozygote individuals in the extreme tail of the PRS distribution characterized by even larger ORs and cumulative incidences than the ones observed in the current analysis.

In addition, our results showed that the inclusion of family history can further and independently improve the risk stratification along with genetic factors. Previous studies have discussed that family history is mainly associated with monogenic variants and minimally with PRS.^{34,35} However, the PRS predictions are affected by estimation errors in variant effect sizes from the reference GWAS; thus, it can be expected that more accurate PRS models will be developed with the increased availability of population-based data.³⁶ Moreover, the additional effect of family history can be caused by unconsidered variants in the genetic risk models (eg, copy number variations), but it can also capture nongenetic contributors such as environmental/lifestyle factors.

Our study has different limitations. First, there is evidence of a healthy volunteers selection bias of the UKB cohort, and thus, the results might not be generalizable in terms of effect sizes.³⁷ Second, our risk assessment was based solely on genetic variants and family history and did not include other risk factors. Previous studies with UKB showed that lifestyle modifiable risk factors play a pivotal role in cancer prevalence,³⁸ and a shared lifestyle within families could influence family history with the disease.³⁹ This might explain the additional effect of family history

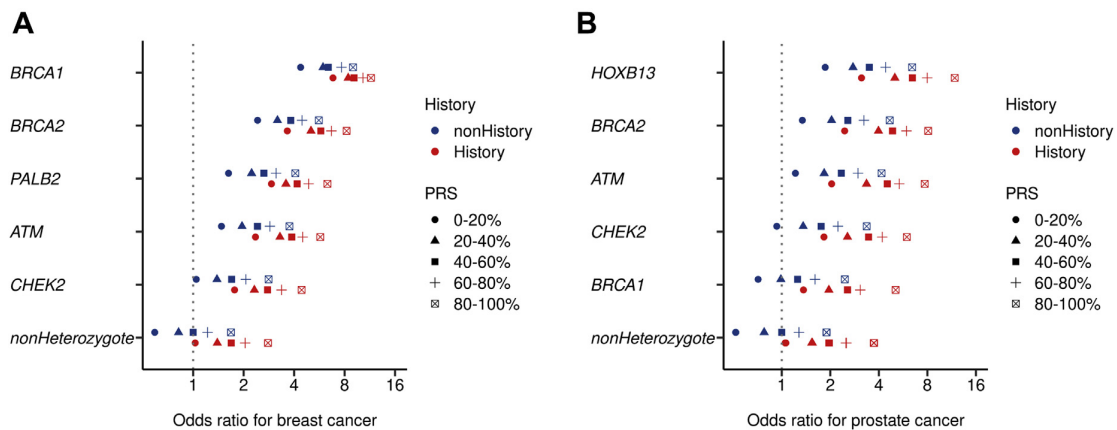


Figure 3 Interplay of pathogenic variant, family history, and PRS in single genes. Odds ratios for cancer were estimated from logistic models adjusted for age and first 4 ancestry principal components for breast cancer (A), and prostate cancer (B). Nonheterozygotes with 40% to 60% PRS and no family history served as the reference group. PRS, polygenic risk score.

of cancer with respect to the genetic risk. Finally, although we performed the analysis on the whole UKB cohort, we could not test the risk stratification generalizability across different populations because of the limited sample size. PRS could be biased toward the European population because PRS was constructed on the basis of European reference GWAS. Thus, PRS might be a worse predictor in non-European or admixed individuals, as previously discussed in different studies.⁴⁰

In conclusion, we showed the significant role of PRS in both general population and heterozygotes of rare pathogenic germline variants in moderate to high-risk cancer genes. PRS strongly alters the penetrance of moderate-risk and high-risk variants and influences the lifetime disease risk. The data suggest that stratification of individuals based solely on the PRS can reach ORs comparable with those associated with heterozygotes of PV in moderate-risk genes that are currently subject to risk-adapted tailored surveillance programs. Consequently, PRS can identify a relatively large group of individuals within the general population for whom intense surveillance measures such as those offered to heterozygotes of moderate-risk genes should be considered. These findings highlight the potential usefulness of PRS in the context of cancer risk stratification. Our analysis shows that family history along with rare PV and PRS represents an additional stratification level to the cancer risk.

Data Availability

Genome-wide genotyping data, exome-sequencing data, and phenotypic data from the UK Biobank are available upon successful project application (<http://www.ukbiobank.ac.uk/about-biobank-uk/>).

Restrictions apply to the availability of these data, which were used under license for the current study (Project ID:

52446). Summary statistics are available from the Polygenic Score Catalog (pgs-info@ebi.ac.uk): for breast cancer at <https://www.pgscatalog.org/score/PGS000007/> and for prostate cancer at <https://www.pgscatalog.org/score/PGS000049/>.

Acknowledgments

This research was conducted using the UK Biobank Resource under application number 52446. C.M. and E.H. were supported by the BONFOR-program of the Medical Faculty, University of Bonn (O-147.0002). P.M. was supported by the Luxembourg National Research Fund as part of the National Centre of Excellence in Research on Parkinson's disease (NCER-PD, FNR11264123) and the Deutsche Forschungsgemeinschaft Research Units FOR2715 (INTER/DFG/17/11583046) and FOR2488 (INTER/DFG/19/14429377). S.A. and I.S. were members of the European Reference Network on Genetic Tumor Risk Syndromes (ERN GENTURIS) Project ID No 739547. The authors acknowledge the use of de.NBI cloud and the support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the Federal Ministry of Education and Research (BMBF) through grant no 031A535A.

Author Information

Conceptualization: E.H., D.R.B., C.M.; Analysis: E.H.; Supervision: P.K., P.M., D.R.B., C.M.; Results Interpretation: A.J.F., I.S., S.A., M.M.N., P.K.; Writing-review and editing: E.H., P.M., R.A., I.S., S.A., M.M.N., P.K., D.R.B., C.M.

Ethics Declaration

Ethics approval for the UK Biobank (UKB) study was obtained from the North West Multicentre for Research Ethics Committee (MREC). The UKB ethics statement is available at <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics>. All UKB participants provided informed consent at recruitment.

Conflict of Interest

The authors declare no conflicts of interest.

Additional Information

The online version of this article (<https://doi.org/10.1016/j.gim.2021.11.009>) contains supplementary material, which is available to authorized users.

References

- Möller S, Mucci LA, Harris JR, et al. The heritability of breast cancer among women in the Nordic twin study of cancer. *Cancer Epidemiol Biomarkers Prev.* 2016;25(1):145–150. <http://doi.org/10.1158/1055-9965.EPI-15-0913>.
- Hjelmberg JB, Scheike T, Holst K, et al. The heritability of prostate cancer in the Nordic twin study of cancer. *Cancer Epidemiol Biomarkers Prev.* 2014;23(11):2303–2310. <http://doi.org/10.1158/1055-9965.EPI-13-0568>.
- Economopoulou P, Dimitriadis G, Psyrri A. Beyond BRCA: new hereditary breast cancer susceptibility genes. *Cancer Treat Rev.* 2015;41(1):1–8. <http://doi.org/10.1016/j.ctrv.2014.10.008>.
- Potter SR, Partin AW. Hereditary and familial prostate cancer: biologic aggressiveness and recurrence. *Rev Urol.* 2000;2(1):35–36.
- Nicolosi P, Ledet E, Yang S, et al. Prevalence of germline variants in prostate cancer and implications for current genetic testing guidelines. *JAMA Oncol.* 2019;5(4):523–528. <http://doi.org/10.1001/jamaoncol.2018.6760>.
- Sakoda LC, Jorgenson E, Witte JS. Turning of COGS moves forward findings for hormonally mediated cancers. *Nat Genet.* 2013;45(4):345–348. <http://doi.org/10.1038/ng.2587>.
- Beebe-Dimmer JL, Yee C, Cote ML, et al. Familial clustering of breast and prostate cancer and risk of postmenopausal breast cancer in the Women's Health Initiative Study. *Cancer.* 2015;121(8):1265–1272. <http://doi.org/10.1002/ncr.29075>.
- Ren ZJ, Cao DH, Zhang Q, et al. First-degree family history of breast cancer is associated with prostate cancer risk: a systematic review and meta-analysis. *BMC Cancer.* 2019;19(1):871. <http://doi.org/10.1186/s12885-019-6055-9>.
- Kuchenbaecker KB, Hopper JL, Barnes DR, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA.* 2017;317(23):2402–2416. <http://doi.org/10.1001/jama.2017.7112>.
- Easton DF, Pharoah PD, Antoniou AC, et al. Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med.* 2015;372(23):2243–2257. <http://doi.org/10.1056/NEJMsr1501341>.
- Mavaddat N, Pharoah PD, Michailidou K, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst.* 2015;107(5):1–15. <http://doi.org/10.1093/jnci/djv036>.
- Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* 2018;50(9):1219–1224. <http://doi.org/10.1038/s41588-018-0183-z>.
- Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet Med.* 2019;21(8):1708–1718. Published correction appears in *Genet Med.* 2019;21(6):1462. <https://doi.org/10.1038/s41436-018-0406-9>.
- Mars N, Widén E, Kerminen S, et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat Commun.* 2020;11(1):6383. <http://doi.org/10.1038/s41467-020-19966-5>.
- Fahed AC, Wang M, Homburger JR, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun.* 2020;11(1):3635. <http://doi.org/10.1038/s41467-020-17374-3>.
- Giri VN, Hegarty SE, Hyatt C, et al. Germline genetic testing for inherited prostate cancer in practice: implications for genetic testing, precision therapy, and cascade testing. *Prostate.* 2019;79(4):333–339. <http://doi.org/10.1002/pros.23739>.
- Pritchard CC, Mateo J, Walsh MF, et al. Inherited DNA-repair gene mutations in men with metastatic prostate cancer. *N Engl J Med.* 2016;375(5):443–453. <http://doi.org/10.1056/NEJMoa1603144>.
- Ewing CM, Ray AM, Lange EM, et al. Germline mutations in HOXB13 and prostate-cancer risk. *N Engl J Med.* 2012;366(2):141–149. <http://doi.org/10.1056/NEJMoa1110000>.
- Sipeky C, Talala KM, Tammela TLJ, Taari K, Auvinen A, Schleutker J. Prostate cancer risk prediction using a polygenic risk score. *Sci Rep.* 2020;10(1):17075. <http://doi.org/10.1038/s41598-020-74172-z>.
- Karlsson R, Aly M, Clements M, et al. A population-based assessment of germline HOXB13 G84E mutation and prostate cancer risk. *Eur Urol.* 2014;65(1):169–176. <http://doi.org/10.1016/j.eururo.2012.07.027>.
- Lecarpentier J, Silvestri V, Kuchenbaecker KB, et al. Prediction of breast and prostate cancer risks in male BRCA1 and BRCA2 mutation carriers using polygenic risk scores. *J Clin Oncol.* 2017;35(20):2240–2250. <http://doi.org/10.1200/JCO.2016.69.4935>.
- Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726):203–209. <http://doi.org/10.1038/s41586-018-0579-z>.
- Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc.* 2015;10(10):1556–1566. <http://doi.org/10.1038/nprot.2015.105>.
- Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434–443. Published correction appears in *Nature.* 2021;590(7846):E53. <https://doi.org/10.1038/s41586-020-2308-7>.
- Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;42(Database issue):D980–D985. <http://doi.org/10.1093/nar/gkt1113>.
- Mavaddat N, Michailidou K, Dennis J, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet.* 2019;104(1):21–34. <http://doi.org/10.1016/j.ajhg.2018.11.002>.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7. <http://doi.org/10.1186/s13742-015-0047-8>.
- Khera AV, Chaffin M, Zekavat SM, et al. Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation.* 2019;139(13):1593–1602. <http://doi.org/10.1161/CIRCULATIONAHA.118.035658>.
- Wray NR, Maier R. Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability. *Curr Epidemiol Rep.* 2014;1(4):220–227. <http://doi.org/10.1007/s40471-014-0023-3>.

30. Gao C, Polley EC, Hart SN, et al. Risk of breast cancer among carriers of pathogenic variants in breast cancer predisposition genes varies by polygenic risk score. *J Clin Oncol*. 2021;39(23):2564–2573. <http://doi.org/10.1200/JCO.20.01992>.
31. Barnes DR, Silvestri V, Leslie G, et al. Breast and prostate cancer risks for male BRCA1 and BRCA2 pathogenic variant carriers using polygenic risk scores. *J Natl Cancer Inst*. Published online July 28, 2021. <https://doi.org/10.1093/jnci/djab147>.
32. Schröder FH, Hugosson J, Roobol MJ, et al. Screening and prostate-cancer mortality in a randomized European study. *N Engl J Med*. 2009;360(13):1320–1328. <http://doi.org/10.1056/NEJMoa0810084>.
33. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer*. 2013;108(11):2205–2240. <http://doi.org/10.1038/bjc.2013.177>.
34. Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. *Ann Oncol*. 2015;26(7):1291–1299. <http://doi.org/10.1093/annonc/mdv022>.
35. Brandão A, Paulo P, Teixeira MR. Hereditary predisposition to prostate cancer: from genetics to clinical implications. *Int J Mol Sci*. 2020;21(14):5036. <http://doi.org/10.3390/ijms21145036>.
36. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759–2772. <http://doi.org/10.1038/s41596-020-0353-1>.
37. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of socio-demographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol*. 2017;186(9):1026–1034. <http://doi.org/10.1093/aje/kwx246>.
38. Kachuri L, Graff RE, Smith-Byrne K, et al. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat Commun*. 2020;11(1):6084. <http://doi.org/10.1038/s41467-020-19600-4>.
39. Al Ajmi K, Lophatananon A, Mekli K, Ollier W, Muir KR. Association of nongenetic factors with breast cancer risk in genetically predisposed groups of women in the UK Biobank cohort. *JAMA Netw Open*. 2020;3(4):e203760. <http://doi.org/10.1001/jamanetworkopen.2020.3760>.
40. Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J. Genetic disease risks can be misestimated across global populations. *Genome Biol*. 2018;19(1):179. <http://doi.org/10.1186/s13059-018-1561-7>.

3.2 Publication 2: Clinically relevant combined effect of polygenic background, rare pathogenic germline variants, and family history on colorectal cancer incidence


Contribution: Data analysis, interpretation of results, writing and revision of manuscript

RESEARCH ARTICLE

Open Access



Clinically relevant combined effect of polygenic background, rare pathogenic germline variants, and family history on colorectal cancer incidence

Emadeldin Hassanin^{1,2†}, Isabel Spier^{3,4,5†}, Dheeraj R. Bobbili², Rana Aldisi¹, Hannah Klinkhammer^{1,6}, Friederike David³, Nuria Dueñas^{7,8}, Robert Hüneburg^{4,9}, Claudia Perne^{3,4}, Joan Brunet^{5,7,8,10}, Gabriel Capella^{5,7,8}, Markus M. Nöthen³, Andreas J. Forstner^{3,11,12}, Andreas Mayr⁶, Peter Krawitz¹, Patrick May², Stefan Aretz^{3,4,5*†}  and Carlo Maj^{1†}

Abstract

Background and aims Summarised in polygenic risk scores (PRS), the effect of common, low penetrant genetic variants associated with colorectal cancer (CRC), can be used for risk stratification.

Methods To assess the combined impact of the PRS and other main factors on CRC risk, 163,516 individuals from the UK Biobank were stratified as follows: 1. carriers status for germline pathogenic variants (PV) in CRC susceptibility genes (*APC*, *MLH1*, *MSH2*, *MSH6*, *PMS2*), 2. low (< 20%), intermediate (20–80%), or high PRS (> 80%), and 3. family history (FH) of CRC. Multivariable logistic regression and Cox proportional hazards models were applied to compare odds ratios and to compute the lifetime incidence, respectively.

Results Depending on the PRS, the CRC lifetime incidence for non-carriers ranges between 6 and 22%, compared to 40% and 74% for carriers. A suspicious FH is associated with a further increase of the cumulative incidence reaching 26% for non-carriers and 98% for carriers. In non-carriers without FH, but high PRS, the CRC risk is doubled, whereas a low PRS even in the context of a FH results in a decreased risk. The full model including PRS, carrier status, and FH improved the area under the curve in risk prediction (0.704).

Conclusion The findings demonstrate that CRC risks are strongly influenced by the PRS for both a sporadic and monogenic background. FH, PV, and common variants complementary contribute to CRC risk. The implementation of PRS in routine care will likely improve personalized risk stratification, which will in turn guide tailored preventive surveillance strategies in high, intermediate, and low risk groups.

Keywords Colorectal cancer, Family history, Hereditary cancer, Polygenic risk, Risk stratification

[†]Emadeldin Hassanin, Isabel Spier, Stefan Aretz and Carlo Maj contributed equally to this work

*Correspondence:

Stefan Aretz

Stefan.Aretz@uni-bonn.de

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Colorectal cancer (CRC) is the fourth leading cancer-related cause of death worldwide. Major established exogenous risk factors are summarized as Western lifestyle [1]. However, an inherited disposition contributes significantly to the disease burden since up to 35% of interindividual variability in CRC risk has been attributed to genetic factors [2, 3].

Around 5% of CRC occur on the basis of a monogenic, Mendelian condition (hereditary CRC), in particular Lynch syndrome (LS) and various gastrointestinal polyposis syndromes. Here, predisposing rare, high-penetrance pathogenic variants (PV, constitutional/germline variants) result in a considerable cumulative lifetime risk of CRC and a syndrome-specific spectrum of extracolonic tumors. The autosomal dominant inherited LS is by far the most frequent type of hereditary CRC with an estimated carrier frequency in the general population of 1:300–1:500 [4–6]. It is caused by a heterozygous germline PV in either of the mismatch repair (MMR) genes *MLH1*, *MSH2*, *MSH6* or *PMS2* or, in few cases, by a large germline deletion of the *EPCAM* gene upstream of *MSH2*. The most frequent Mendelian polyposis syndrome is the autosomal dominant Familial Adenomatous Polyposis (FAP) caused by heterozygous germline PV in the tumor suppressor gene *APC*, followed by the autosomal recessive *MUTYH*-associated polyposis (MAP) which is based on biallelic germline PV of the base excision repair gene *MUTYH* [7, 8]. However, even in such monogenic conditions, the inter- and intrafamilial penetrance and phenotypic variability is striking, pointing to modifying exogenous or endogenous factors. Heterozygous (monoallelic) *MUTYH* germline PV may be associated with a slightly increased CRC risk [9, 10]; the carrier frequency in northern European populations is estimated to be 1:50–1:100 [4].

Approximately 20–30% of CRC cases are characterized by a suspicious, but unspecific familial clustering of CRC (familial CRC). Around 25% of CRC cases occur before 50 years of age (early-onset CRC); in around one quarter of those a hereditary type (mainly LS) has been identified [11]. Although further high-penetrance candidate genes have been proposed [12–14], the majority of familial and early-onset cases cannot be explained by monogenic subtypes and instead are supposed to result from a multifactorial/polygenic etiology including several moderate-/intermediate penetrance risk variants and shared environmental/lifestyle factors. A positive family history (FH) in first- and second-degree relatives increases the risk of developing CRC by 2- to ninefold [15, 16], which underpins the hypothesis of shared genetic and non-genetic risk factors.

A variety of models to predict CRC risk has been developed and evaluated, which include clinical data, FH, lifestyle factors, and genetic information [17]. For more than a decade, genome-wide association studies (GWAS) in large unselected CRC cohorts identified an increasing number of common, low-penetrance risk variants, mainly single nucleotide polymorphisms (SNPs), which are significantly associated with CRC risk [18–21]. Each SNP risk allele individually contributes only little to CRC risk (OR 1.05 to 1.5), however, summarised in quantitative polygenic risk scores (PRS), the combined effect might explain a substantial fraction of CRC risk variability and can identify individuals at several times lower and greater risk than the general population [22–24].

As such, it is expected that the genetic background defined by the common risk variants may not only influence the occurrence of late-onset sporadic cases, but also modulate the risk of familial, early-onset, and hereditary CRC [25]. Recent studies demonstrated that high PRS values are associated with an increased risk of CRC and other common cancers in the general population up to an order of magnitude that is almost similar to hereditary tumor syndromes [26, 27].

Based on these data, it can be hypothesized, that the identification of common genetic CRC risk variants not only provides deep insights into the biological mechanisms and pathways of tumorigenesis, but could improve personalized risk stratification for sporadic, familial/early-onset, and hereditary CRC in the future by the implementation of SNP-based PRS screening in routine patient care, which will in turn guide tailored preventive strategies in high, moderate, and low risk groups.

However, even if previous studies provide promising results for a clinical benefit of a PRS-based personalized risk stratification, the impact of common risk factors and their interplay with high-penetrance variants and other unspecified factors, captured partly by the FH, still has to be improved and validated in additional patient cohorts.

In the present work, we compare the prevalence and the lifetime risk of CRC among 163,516 individuals from a population-based European repository (UK Biobank, UKBB). Individuals were stratified according to three major risk factors 1) their carrier status of rare, high-penetrance.

Methods

Data source

UK Biobank (UKBB) genetic and phenotypic data were used in this study. UKBB is a long-term prospective population-based cohort study that has recruited volunteers mostly from England, Scotland, and Wales, with over 500,000 participants aged 40 to 69 years at the time of recruitment. For each participant, extensive phenotypic

and health-related data is available; genotyping data is accessible for 487,410 samples, and exome sequencing data is available for 200,643 people. All participants gave written consent, and the dataset is available for research. UKBB provided follow-up information by linking health and medical records [28].

Study participants

CRC cases were defined based on self-reported code of 1022 or 1023 (in data field 20,001), or ICD-10 code of C18.X or C20.X, D01.[0,1,2], D37.[4, 5], or ICD-9 of 153.X or 154.[0,1] (in hospitalization records). Control samples were those that had no previous diagnosis of any cancer. The study includes people of all ethnicities. Outliers for heterozygosity or genotype missing rates, putative sex chromosome aneuploidy, and discordant reported sex versus genotypic sex were excluded. Only individuals ($n=200,643$) who had both genotyping and whole-exome sequencing (WES) data were considered. If the genetic relationship between individuals was closer than the second degree, defined as kinship coefficient >0.0884 as computed by the UK Biobank, we removed one from each pair of related individuals (cases were retained if exist).

Variant selection

We used ANNOVAR [29] to annotate the VCF files from the 200,643 WES samples. The Genome Aggregation Database (gnomAD) [30] were used to retrieve variant frequencies from the general population. We focused on rare PV for hereditary CRC (Lynch syndrome, polyposis) and considered the same variant filtering approach that was used in a recent study aiming at selecting rare PV [31]. The following inclusion criteria were used: (1) only *APC*, *MUTYH*, *MLH1*, *MSH2*, *MSH6*, *PMS2* variants in protein-coding regions were included since PV in other genes associated with hereditary CRC are too rare or even absent in the study population; (2) allele frequency (AF) <0.005 in at least one ethnic subpopulation of gnomAD; (3) not annotated as “synonymous,” “non-frameshift deletion” and “non-frameshift insertion”; (4) annotated as “pathogenic” or “likely pathogenic” based on ClinVar [32]. We did not include *MUTYH* in the pooled analysis since no biallelic (i.e. high penetrance) case was identified in the cohort; however, we included the heterozygous (monoallelic) carriers in the single gene analysis to compare the effect size with the other genes.

Polygenic risk scores (PRS)

We applied a previously validated PRS for CRC with 95 variants to calculate the PRS [18]. The PRS was estimated using the PLINK 2.0 [33] scoring function through UKB genotype data. To reduce PRS

distributions variance among genetic ancestries, we used a previous approach [34]. We used the first four ancestry principal components (PCs) to fit a linear regression model to predict the PRS across the full dataset ($pPRS \sim PC1 + PC2 + PC3 + PC4$). Adjusted PRS (aPRS) were calculated by subtracting pPRS from the raw PRS and used for the subsequent analysis.

In addition, we calculated the PRS using 140 SNPs [18] and another PRS based on 50 SNPs that were replicated in the meta-analyzed GWAS after excluding UKBB samples [35]. Thus, in total three PRS models were computed: (1) 95 SNPs (95 PRS); (2) 140 SNPs (140 PRS); (3) 50 SNPs (50 PRS).

Statistical analysis

Individuals were divided into groups depending on (1) carrier status of PV, (2) PRS, and (3) FH. For FH, we considered participants' reports of CRC in their parents and siblings (data fields: 20,110, 20,107, 20,111). For PRS, individuals were assigned into three groups: low ($<20\%$ PRS), intermediate (20–80% PRS), and high ($>80\%$ PRS) where the definition of a high PRS (above the 80th percentile) corresponding to $OR \geq 2$.

We conducted both an analysis specific to single genes and a combined analysis (i.e., carriers of PV in *APC*, *MLH1*, *MSH2*, *MSH6* and *PMS2*). First, we estimated the OR for each carrier group based on a logistic regression adjusting for age at recruitment, sex, CRC screening status, and the first four ancestry PCs. Afterwards, we additionally incorporated interactions between PV carriers and FH with PRS by introducing an interaction term within the logistic regression model.

We calculated the lifetime risk by age 75 from carrier status of rare PV and the PRS and hazard ratios (HRs) based on a Cox proportional hazards model. Individual's age served as the time scale, representing the time to event, for observed cases (age at diagnosis), and censored controls (age at last visit); age 0 was used as index time. Carrier status, PRS category, FH, age, sex, CRC screening status, and the first four ancestry PCs were incorporated in the model, and adjusted survival curves were produced. The information about FH and CRC screening is based on interviews at the time of study recruitment. However, information about the timing or result of CRC screening was not available. We therefore included this information as a binary covariate to account for both effects.

Model performance was assessed via the area under the receiver operating characteristic curve (AUC), Nagelkerke's Pseudo- R^2 , and the C-index for time-to-event data. R 3.6.3 with the corresponding add-on packages *survival* and *survminer* was used for all statistical analyses.

Results

Stratification of UKBB individuals for CRC prevalence, FH, and PV carrier status

We identified 1,902 CRC cases (894 prevalent cases and 1,008 incident cases) among the 163,516 UKBB individuals that retained after exclusion criteria, with a mean age at diagnosis of 60.9 years. The remaining 161,614 individuals with no previous diagnosis of any cancer were considered as controls, with a mean age of 56.9 years at last visit (Table 1). The European population represents 92% of the analyzed cohort.

The fraction of individuals with a positive FH of CRC is significantly higher in cases (19%) compared to controls (11%) (OR = 1.95 [1.73–2.19], $P < 0.01$) and ranges between 9 and 23% in the subgroups (Table 2). There is a significantly higher proportion of individuals with a FH of CRC not only among carriers of PV in the selected cancer susceptibility genes (OR = 1.96 [1.72–2.20], $P < 0.01$), but also among non-carriers with high PRS (OR = 1.60 [1.31–1.94], $P < 0.01$).

In the analyzed CRC susceptibility genes *APC*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, we identified 399 heterozygous carriers of 111 PV. They were present in 30 (1.57%) cases and 369 (0.23%) controls, which is in line with published data. A list of the considered variants and annotations is shown in Additional file 1: Table S1, a summary of the number of PV carriers per gene is provided in

Table 1 Characteristics of the 163,516 UK Biobank participants by colorectal cancer (CRC) status

	Cases	Controls
Participants, n	1902	161,614
Male, n (%)	1017 (53.47)	73,979 (45.78)
Female, n (%)	885 (46.53)	87,635 (54.22)
Age, mean (SD)	60.96 (8.56)	56.91 (8.51)
Carriers, n (%)	30 (1.58)	369 (0.23)
Family history of colorectal cancer, n (%)	368 (19.35)	17,696 (10.95)

Table 2 Characteristics of the UK Biobank participants by carrier status and polygenic risk score (PRS) strata

Carrier status	Carrier			Non carrier		
	High	Intermediate	Low	High	Intermediate	Low
Participants, n	74	247	78	32,628	97,863	32,626
Cases, n (%)	11 (14.86)	16 (6.48)	3 (3.85)	686 (2.1)	1004 (1.03)	182 (0.56)
Controls, n (%)	63 (85.14)	231 (93.52)	75 (96.15)	31,942 (97.9)	96,859 (98.97)	32,444 (99.44)
Male, n (%)	36 (48.65)	110 (44.53)	35 (44.87)	14,824 (45.43)	45,115 (46.1)	14,876 (45.6)
Female, n (%)	38 (51.35)	137 (55.47)	43 (55.13)	17,804 (54.57)	52,748 (53.9)	17,750 (54.4)
Age at assessment, mean (SD)	57.12 (8.89)	56.16 (9.15)	57.35 (8.42)	56.93 (8.52)	56.97 (8.51)	56.96 (8.53)
Family history of colorectal cancer, n (%)	17 (22.97)	53 (21.46)	18 (23.08)	4300 (13.18)	10,671 (10.9)	3005 (9.21)

Additional file 2: Table S2. No individual with a homozygous PV was identified. In other known genes associated with hereditary CRC (*BMPRIA*, *POLE*, *POLD1*, *RNF43*, *SMAD4*, *STK11*), the number of (L)P variant carriers was extremely low or no variant carrier was present at all, so that these genes were not considered in the analysis.

PRS distribution within the UKBB cohort

CRC PRS follow a normal distribution both regarding raw and PC-adjusted PRS (Additional file 2: Fig. S1) and is significantly higher in cases compared to controls, regardless of which PRS model is used (95 PRS, 140 PRS or 50 PRS), the PRS is significantly higher in cases compared to controls (Additional file 2: Fig. S2). The OR for 50 PRS (1.74 [1.57–1.92]) is slightly lower than that of 95 PRS (1.98 [1.79–2.19]), or 140 PRS (1.92 [1.74–2.12]); that might be due to overfitting.

Since we included only individuals with both genotyping and WES data, we investigated the distribution of the PRS and age in the whole cohort and compared it to the subcohort with WES data. Density plots show that the distribution of PRS and age was similar between both groups (Additional file 2: Fig. S3).

The prevalence of CRC according to PRS percentiles demonstrates that values in the extreme right tail of the PRS distribution are associated with a non-linear increase of CRC risk, whereas in the left tail a less evident non-linear decrease can be observed (Additional file 2: Fig. S4). This supports the hypothesis of using PRS to stratify individuals into risk classes (i.e., low, intermediate, and high risk) according to a liability threshold model.

Interplay between PV and PRS

There was no overlap between the selected rare high penetrance PV and the common SNPs used for PRS calculation, and thus, the PRS represents an additional genetic signal. Notably, the PRS distributions showed that the mean of PRS is significantly higher in affected carriers

compared to unaffected carriers ($P < 0.01$) (Additional file 2: Fig. S5).

We assessed how CRC risk is influenced by PRS and carrier status for PV in high penetrant CRC susceptibility genes (*APC*, *MLH1*, *MSH2*, *MSH6*, *PMS2*) by calculating the ORs for CRC across groups compared to non-carriers with intermediate PRS as reference group. Non-carriers with a low or high PRS are estimated to have a 0.5-fold or 2.1-fold change in the odds for CRC, respectively. We observed that the PRS also alters the penetrance of PV in susceptibility genes considerably as PV carriers with high PRS had four times higher OR than carriers with low PRS

(OR = 17.5 and 3.9, respectively; Fig. 1A; and corresponding HR in Additional file 2: Table S3). We did not observe a significant interaction between PV carrier status and PRS ($p = 0.87$). In addition, we performed a sensitivity analysis including only the incident cases ($n = 1,008$). We observed the same trend, that PRS provides an OR risk gradient in the general population and among carriers of pathogenic variants in CRC susceptibility genes (Additional file 2: Table S4).

The high PRS, which is by definition present in 20% of the non-carriers, is associated with an almost doubled CRC risk (Fig. 1A, Table 2). Since the vast majority

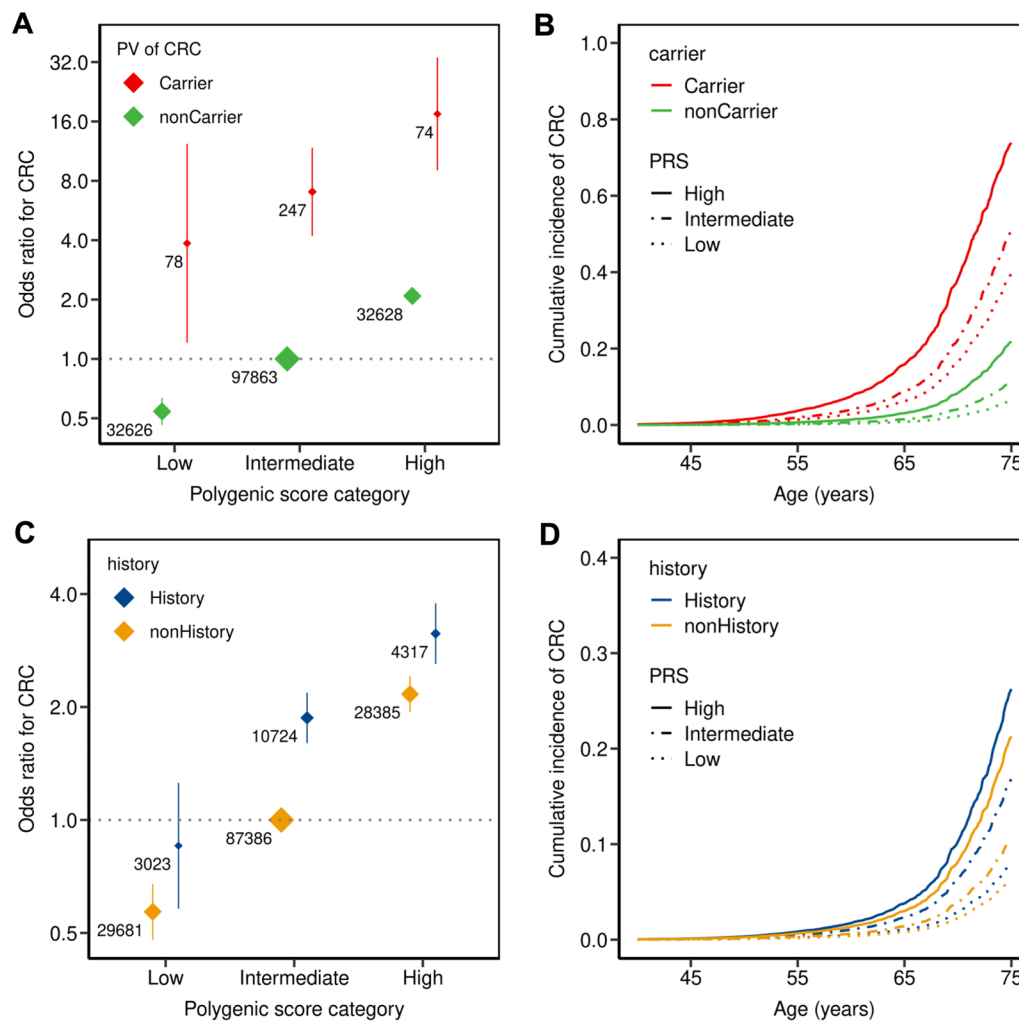


Fig. 1 Colorectal cancer odds ratio and cumulative incidence stratified by carrier and family history status. Individuals stratified for PV carrier status (A + B), and family history (first-degree relative with CRC) (C + D) into three strata based on their polygenic risk score (PRS): Low (< 20% percentile), intermediate (20–80% percentile), or high (> 80% percentile) PRS. The odds ratio (OR) was calculated from a logistic regression model with age, sex, CRC screening status, and the first four principal components of ancestry as covariates. The reference group was non-carriers with intermediate PRS (A), and no family history with intermediate PRS (C). The adjusted OR is indicated by the colored boxes. The numbers next to the ORs indicate the sample size of the corresponding group. The 95% confidence intervals are indicated by the vertical lines around the boxes. Cumulative incidence was estimated from a cox-proportional hazard model using age, sex, family history, CRC screening status, and the first four ancestry principal components as covariates

(97.9%) of non-carriers are controls (=healthy), almost the same percentage results if only healthy non-carriers are considered. We performed the same analysis using the 140 PRS and 50 PRS. All the three PRS models had comparable performance in the UKBB cohort (Additional file 2: Fig. S6).

Similarly, the lifetime cancer risk analysis shows a combined impact of PV and PRS: Among carriers, the estimated cumulative incidence by age 75 increased from 40% in case of a low PRS to 74% in case of a high PRS compared to 6% to 22% for non-carriers (Fig. 1B, Additional file 2: Table S3).

Inclusion of family history on cancer risk stratification

Taking individuals with no FH and intermediate PRS as a reference, both FH and PRS are associated with a higher CRC risk (Fig. 1C, Additional file 2: Table S5). The CRC risk for individuals having low PRS and no FH (OR 0.6) is five times lower than for individuals having both positive FH and high PRS (OR 3.1). We did not observe a significant interaction between FH status and PRS ($p=0.12$). Noteworthy, individuals without FH and high PRS and individuals with FH and intermediate PRS both have similar CRC risks with an OR of around 2, whereas the CRC risk of individuals having low PRS even in the context of a FH is decreased compared to the reference group.

Among individuals with FH, the cumulative CRC incidence by age 75 increases threefold from 8% in case of a low PRS to 26% in case of a high PRS (Fig. 1D). Noteworthy, the cumulative CRC incidence of individuals with a positive FH and an intermediate PRS is lower (16%) than for individuals with negative FH and a higher PRS category (21%), respectively.

The full model integrating PRS, FH, and PV status shows that the CRC risk is strongly influenced by PRS in all groups (Fig. 2, Additional file 2: Table S6). Considering the non-carriers with no FH and intermediate PRS group as reference, the CRC OR in low PRS is 0.6 for non-carriers with no FH, while it is estimated more than 60 times higher (OR 40) for carriers with FH and high PRS (Fig. 2A). The corresponding cumulative CRC incidences are 6% and 98%, respectively (Fig. 2B). Although all PV carriers showed a significantly increased CRC risk, both the PRS and FH modify these risks considerably: depending on the FH and PRS, the OR in PV carriers vary between 4 and 40 and the cumulative incidence between 35 and 98%. Despite the CRC screening status is a key predictor for CRC risk (Additional file 2: Fig. S7), the main findings of the analysis were maintained irrespective of the screening status.

PRS improved model discrimination over carrier status and FH of CRC in first-degree relatives. The AUC derived from PRS (0.688) was higher compared to those derived using FH (0.654) and carrier status (0.646). The full model including PRS, carrier status, and FH improved the AUC (0.704) in risk prediction by 1.6%, 5%, and 5.8%, respectively, and was also better than any combination of two factors (Table 3, Additional file 2: Fig. S8a). We also performed an analysis in which age and sex were excluded. The AUCs demonstrate that the PRS still has a high discriminative power for CRC risk prediction (Additional file 2: Fig. S8b).

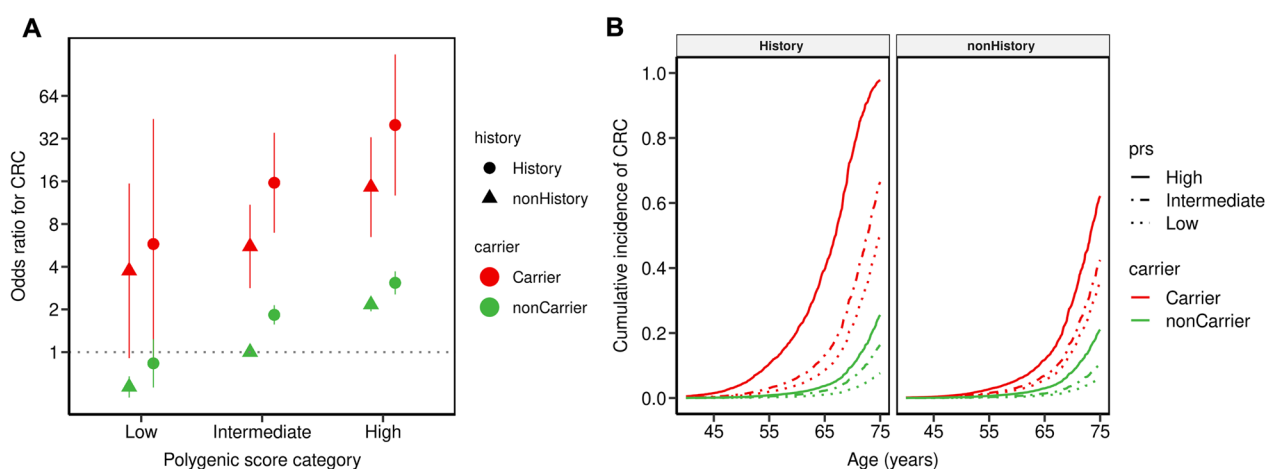
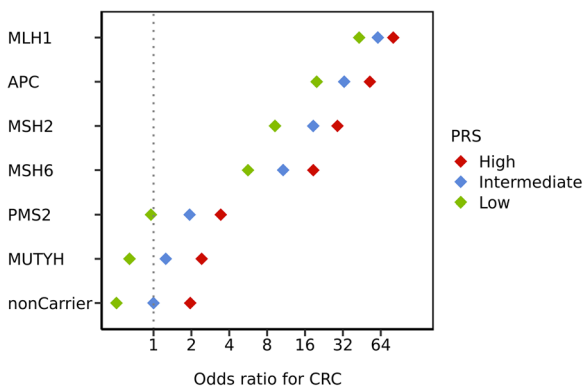


Fig. 2 Interplay of pathogenic variant carrier status, family history, and polygenic risk score. **A** Colorectal cancer (CRC) odds ratios (ORs) were estimated from logistic models adjusted for age, sex, CRC screening status, and first four ancestry principal components. Non-carriers with intermediate PRS and no family history served as the reference group. **B** Cumulative incidence was estimated from a cox-proportional hazard model using age, sex, family history, CRC screening status and the first four ancestry principal components as covariates

Table 3 Model discrimination assessed for combinations of polygenic risk score, family history of CRC and carrier

	AUC (C.I 95%)	C-index	Nagelkerke's Pseudo-R ²
PRS + FH + carrier	0.704 (0.68–0.73)	0.657	0.055
PRS + FH	0.698 (0.67–0.72)	0.652	0.052
PRS + carrier	0.693 (0.66–0.71)	0.646	0.051
PRS	0.688 (0.66–0.71)	0.640	0.049
FH + carrier	0.660 (0.64–0.68)	0.580	0.032
FH	0.654 (0.63–0.68)	0.574	0.031
Carrier	0.646 (0.62–0.67)	0.556	0.030

**Fig. 3** Interplay of pathogenic variant carrier status, family history, and polygenic risk score in single genes. Odds ratios (ORs) for colorectal cancer (CRC) were estimated from logistic models adjusted for age, sex, CRC screening status, and first four ancestry principal components. Non-carriers with intermediate PRS and no family history served as the reference group

The impact of polygenic risk in single gene mutation carriers

The gene-specific analysis revealed a strong variability in risk conferred by rare heterozygous PV in the different genes. The largest effect sizes are attributable for *MLH1* and *APC*, those for *MSH2* and *MSH6* are a bit less, while the effect size for *PMS2* is considerably lower (Fig. 3). When heterozygous *MUTYH* variants are included in this analysis, the risks are very similar to the *PMS2*-related risks. Both the *PMS2* and heterozygous *MUTYH* risks show a broad overlap with the non-carrier risks, while there is no overlap between the risks of non-carriers and those with PVs in *MLH1*, *MSH2*, *MSH6*, and *APC*.

We estimated how PRS and FH influence CRC prevalence among PV carriers in each of the five susceptibility genes (Additional file 2: Table S7). Despite the different effect sizes, the PRS and FH modifies the relative risk across all genes; however, the effect of PRS and

FH is conversely related to the penetrance of the gene with the smallest effects in *MLH1* PV carriers.

As for the overall analysis, in the gene-specific analysis a positive FH, a PV in a cancer risk gene, and a high PRS are associated with an increased CRC risk. As such, an individual with a low-penetrance *PMS2* PV, but high PRS and/or positive FH ends up with an estimated CRC risk similar to a *MSH6* PV carrier without FH and/or low PRS (Additional file 2: Fig. S9, Table S7).

Discussion

Recent studies demonstrated that the polygenic background, defined as PRS based on disease-associated SNPs, modifies the risks for several cancers of the general population including CRC considerably, both in terms of age at onset and cumulative lifetime risks [12, 23, 27, 36–38]. In line with this, the risk alleles of those SNPs are found to also accumulate in unexplained familial and early-onset CRC cases [25, 39]. Whereas a low polygenic burden decreases the CRC risk down to one quarter on average, individuals with a high PRS (> 80%) doubles and those with a very high PRS (99%) almost quadruplicate their risk and thus, reach a CRC risk in an order of magnitude almost comparable to carriers of hereditary CRC with low PRS [31]. In a previous study, Jia et al. found that the risk of CRC is significantly associated with its PRS: Compared with individuals in the lowest PRS quintile those in the highest quintile had a greater than three-fold risk (during a 5.8-year follow-up period). Hazard Ratios estimated with the middle quintile as the reference resulted in a risk between 0.56 and 1.71, a threefold risk in those in the top 1% of PRS, and a 70% reduced CRC risk for individuals in the bottom 1% of the PRS [38].

To extend these studies on how the CRC prevalence is influenced by genetic susceptibility using, we used the sufficiently larger, more robust dataset of the most recent UKBB cohort, incorporate the family history (FH) as an additional factor for risk stratification, and include a single gene analysis. We considered both the genetic component driven by rare high-penetrance PV associated with hereditary CRC and common low-penetrance variants captured by the PRS.

Firstly, our results confirm that the polygenic background strongly modulates CRC risk in the general population. Compared to the average polygenic burden, individuals with a low (< 20%) or high (> 80%) PRS are estimated to have a 0.5-fold or 2.1-fold change in the odds for CRC, respectively. The additional time-to-event analysis revealed a corresponding cumulative lifetime risk of 6% and 22% by age 75. Hence, when the PRS is included in risk calculation, around 20% of healthy individuals of the general population with no FH of CRC have a doubled CRC risk, which is similar

to those with a first degree relative affected by CRC [40]. These so far unknown and otherwise unrecognisable at-risk individuals might need surveillance 10–15 years earlier than usually recommended [41]. On the other hand, the around 20% of individuals with low PRS and no FH might need less surveillance than the general population due to a considerably lowered risk, while even those with low PRS and positive FH might not need a more intense surveillance than the general population.

A concern in evaluating CRC PRS using 95 or 140 SNPs [18] in UKBB studies is that the calculation is based on summary statistics derived from a GWAS meta-analysis that included findings from the UKBB. Previous studies have also used 95 or 140 SNPs, but it is uncertain if this could result in overfitting of models. A recent study [35] addressed this issue using stringent inclusion criteria, only including 50 SNPs that reached GWAS significant ($p < 5 \times 10^{-8}$) in the meta-analysis after excluding UKBB samples. The effect sizes from meta-analysis of these 50 SNPs were then used to conduct the 50 PRS. The slightly lower OR of the 50 PRS in the present study compared to the 95 and 140 PRS might be due to overfitting; however, by comparing the PRS calculations, we could show that all three PRS models had a comparable performance in the UKBB cohort (Additional file 2: Figs. S2 and S6).

It is well known that among patients with hereditary CRC syndromes, the age of onset and cumulative CRC incidence is very heterogeneous, even within PV carriers of the same family. The estimated gene-specific, individual CRC lifetime risks of LS patients with *MLH1* or *MSH2* PV can be lower than 10% but as high as 90–100% in a considerable fraction. In the past, the analysis of modifying effects based on common CRC-associated variants in LS and other high-risk groups has been restricted to selected cohorts and small subsets of SNPs [42, 43]. A recent study demonstrated that the polygenic background also substantially influences the CRC risk in LS using UKBB data, even though the ORs for CRC risks could only be predicted due to the small sample sizes [31]. In the present work, ORs could be calculated directly from the model since over three times more UKBB individuals have been included with six times more CRC cases, and five times more PV carriers.

So secondly, we were able to show that the PRS modifies the CRC risks not only in the general population considerably, but also in carriers of a MMR gene PV identified in the general population. For the first time we demonstrated, that this is also true for *APC* PV. Depending on the PRS, the cumulative CRC lifetime incidence in PV carriers ranged between 40 and 74%, and thus, the PRS is able to explain parts of the interindividual variation in CRC risk among PV carriers.

However, the single-gene analysis revealed heterogeneous effects across genes and therefore the modifying role of the polygenic background should be framed within the absolute risk attributable to individual genes. As expected, the effect of the PRS seems to be relevant in particular in less penetrant CRC risk genes such as *PMS2* where the OR ranges between 0.94 and 5.43 respectively (Additional file 2: Table S6). This is in line with findings in moderate breast cancer risk genes such as *CHEK2*, *PALB2* and *ATM* [44–46] and suggests that PRS inclusion in risk stratification may in particular be relevant to prevent excess of surveillance measures in PV carriers of those genes.

In addition, our results provide evidence that the inclusion of FH can further and independently improve the risk stratification in both carriers and non-carriers. Including PRS and FH in risk assessment, the cumulative CRC lifetime incidence ranged between 8 and 26%, and in PV carriers between 30 and 98%, and thus, outperformed the consideration of a single risk factor. This suggests that familial clustering points to additional risk factors besides those captured by common low-risk SNPs (PRS) and rare PV [47, 48]. These might be common and rare structural genetic alterations including copy number variants, rare non-coding variants, or other intermediate and low-impact risk variants not included routinely in PRS models, and non-genetic contributors such as environmental/lifestyle factors.

Only few PRS studies considered the FH. In line with our results, Jenkins et al. found no correlation between SNP-based and FH-based risks and an improved risk stratification when both PRS and FH are considered [47]. In the analyses by Jia et al., the AUC derived from PRS (0.609) was substantially higher compared to the one derived using FH (0.523). Adding PRS and FH of cancer in first-degree relatives improved the model's discriminatory performance (AUC 0.613) [17, 49]. Our AUC calculations point in the same direction with a higher AUC (0.704) when all three risk factors (PRS, FH, carrier status) are considered.

Interestingly and in apparent contrast to our results and those of others, a study using 826 European-descent carriers of PV in the DNA MMR genes *MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM* (i.e. LS carriers) from the Colon Cancer Family Registry (CCFR) did not find evidence of an association between the PRS and CRC risk, irrespective of sex or mutated gene, although an almost identical set of SNPs was used for PRS calculations [50]. A reason which might partly explain different risk estimates between studies using individuals from a population-based repository such as the UKBB and those using curated clinical data registries, where patients/families with suspected hereditary disease are included (e.g. the

CCFR), is a potentially different risk composition across cohorts recruited in different ways (recruitment bias). That way, a familial clustering of CRC might reflect the existence of several genetic and non-genetic risk factors as outlined above, which are not captured by the PRS and which may superimpose the polygenic impact.

In particular, the composition of cases and controls is different between the Jenkins et al. study on the one hand and the Fahed et al. and present study on the other hand. In the Jenkins et al. study, obviously both cases (i.e., PV carriers with CRC) and controls (healthy PV carriers) derived from the same LS families, while the UKBB controls are PV carriers not apparently related to the PV cases. This is also reflected by the different ratio between cases and controls (7.5% CRC cases among PV carriers in the present study, but 61% in the Jenkins et al. study). Hence, the controls in the Jenkins et al. study are relatives of the cases and thus, it is likely that they share parts of the polygenic background and other risk factors of their affected relatives (cases) to a certain extent which may explain the observed missing effect of the PRS. The comparison between population-based and registry-based predictions indicates that the study design and recruitment strategy may strongly influence the results and conclusions. Consequently, the application of PRS in clinical practice should consider the familial background and ascertainment of the patient.

Our data analyses provide evidence that the PRS acts as a relevant risk modifier for CRC among both the general population and population-based PV carriers in genes causing hereditary CRC. The findings of us and others qualify the PRS as important component of risk stratification and resulting risk-adapted surveillance strategies in terms of age of onset and frequency. Given the risk distribution across PRS groups, the PRS can define a considerable proportion of the general population at a CRC risk level which is considered sufficient for a more or a less intensive surveillance. Importantly, the non-carriers with high PRS are a much larger target group compared to PV carriers and thus might generate an even higher preventive effect from a healthcare perspective. A small group of non-carriers with positive FH and high PRS even has CRC risks almost in the same order of magnitude as LS carriers without additional risk factors and thus may need similar intensive surveillance measures.

According to these findings, there should be a potential benefit for both the general population and at-risk individuals carrying PV, from the inclusion of PRS in healthcare prevention policies, as risk-stratified surveillance improves early disease detection and prevention. A recent study demonstrated that individuals with a higher genetic risk benefited more substantially from preventive measures than those with a lower risk: CRC screening

was associated with a significantly reduced CRC incidence and more than 30% reduced mortality among individuals with a high PRS [51, 52]. Preliminary calculations indicate that polygenic-risk-stratified CRC screening could become cost-effective under certain conditions including an AUC value above 0.65 which was reached in our analyses [53].

Based on the striking different penetrance between individual hereditary CRC genes, very recent guidelines start to recommend a more gene-specific surveillance intensity in LS and polyposis [54, 55]. Given the strong modifying effect, the inclusion of additional risk factors will result in a more appropriate, clinically relevant risk stratification. Our results demonstrate that a combined risk assessment including FH and PRS will likely improve precise risk estimations and tailored preventive measures not only in the general population, but also in patients with hereditary disease.

Our study has some limitations. Firstly, there is evidence of a “healthy volunteers” selection bias of the UKBB population (UKBB participants tend to be healthier than the general population), and thus the results might not be completely generalizable in terms of effect sizes [56]. Secondly, we cannot exclude that few carriers of APC PV who were classified as controls, are affected by a polyposis but have not been recognized as such or did not develop CRC due to intensive surveillance and/or prophylactic surgery, so that the calculated CRC risk of APC PV might be slightly underestimated. As in other similar studies, the presence of colorectal polyps could not be considered due to the lack of appropriate data. Thirdly, to increase the power of the analysis, our risk assessment was based solely on genetic variants and FH and did not include other risk factors. Previous studies on UKBB cohorts showed that lifestyle modifiable risk factors play a pivotal role in cancer prevalence, and a shared lifestyle within families could influence FH with the disease [49, 57]. That might explain the partly independent association of the FH and the genetic risk. Finally, although we performed the analysis on the whole UKBB cohort, we could not test the risk stratification generalizability across different populations due to the limited sample size. PRS could be biased towards the European population as PRS was constructed based on European reference GWAS. Thus, these PRS might be a worse predictor in non-European or admixed individuals, as previously discussed in different studies [58].

Conclusion

In conclusion, we show the important role of PRS and FH on CRC risk in both the general population and population-based carriers of a monogenic predisposition

for CRC. The combined effect of common variants can strongly alter the age-related penetrance and life-time risk of CRC. Thus, the PRS represents an additional, independent stratification level to cancer risk besides the FH and lifestyle factors and likely increase the accuracy of risk estimation. Consequently, PRS can define a relevant proportion within the general population as a risk group, which should be considered as subjects for more intense surveillance measures, and in addition point to a striking risk variability even among carriers of hereditary CRC, which requires more personalized, risk-adapted surveillance strategies. As expected, the modifying effect of the PRS seems to be relevant in particular for moderate penetrant CRC risk genes. When important modifiers such as polygenic background, FH, and non-genetic factors are included in risk assessment, the dichotomous risk division between sporadic and hereditary CRC will be partly replaced by a more continuous risk distribution.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12920-023-01469-z>.

Additional file 1: Table S1. A list of the considered variants and annotations.

Additional file 2. Supplemental Figures 1–9 and Supplemental Tables 1–7.

Acknowledgements

UK Biobank analyses were conducted via application 52446 using a protocol approved by the Partners HealthCare Institutional Review Board. CM and EH are supported by the BONFOR-program of the Medical Faculty, University of Bonn (O-147.0002). This study was supported (not financially) by the European Reference Network on Genetic Tumour Risk Syndromes (ERN GENTURIS)—Project ID No 739547. ERN GENTURIS is partly co-funded by the European Union within the framework of the Third Health Programme “ERN-2016—Framework Partnership Agreement 2017–2021”. DRB and PM are supported by the FNR INTER/DFG/21/16394868. This research was also supported by the Instituto de Salud Carlos III and co-funded by European Social Fund—ESF investing in your future—(grants CM19/00099 and PID2019-111254RB-I00) and from the European Union’s Horizon 2020 research and innovation program under the EJP RD COFUND-EJP N° 825575.

Author contributions

EH performed the statistical analysis and the bioinformatics. EH, IS, DRB, CM and SA conceived and designed the study. EH, IS, CM, and SA drafted the initial manuscript. RA, HK, FD, ND, RH, CP, JB, GC, MMN, AJF, AM, PK, and PM performed the critical expert revision. PK, PM, CM, and SA supervised the study. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. No funding was obtained for this study.

Availability of data and materials

Genome-wide genotyping data, exome-sequencing data, and phenotypic data from the UK Biobank are available upon successful project application (<http://www.ukbiobank.ac.uk/about-biobank-uk/>). Restrictions apply to the availability of these data, which were used under license for the current study

(Project ID: 52,446). Summary statistics are available from the Polygenic Score Catalog (pgs-info@ebi.ac.uk): <https://www.pgscatalog.org/publication/PGP000170/>

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing Interests

No potential conflicts (financial, professional, or personal) relevant to the manuscript.

Author details

¹Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn, Germany. ²Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-Sur-Alzette, Luxembourg. ³Institute of Human Genetics, Medical Faculty, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. ⁴National Center for Hereditary Tumor Syndromes, University Hospital Bonn, Bonn, Germany. ⁵European Reference Network on Genetic Tumour Risk Syndromes (ERN GENTURIS) – Project ID No 739547, Nijmegen, The Netherlands. ⁶Medical Faculty, Institute for Medical Biometry, Informatics and Epidemiology, University Bonn, Bonn, Germany. ⁷Hereditary Cancer Program, Catalan Institute of Oncology-IDIBELL, ONCOBELL, Hospitalet de Llobregat, Barcelona, Spain. ⁸Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Instituto Salud Carlos III, Madrid, Spain. ⁹Department of Internal Medicine I, University Hospital Bonn, Bonn, Germany. ¹⁰Hereditary Cancer Program, Catalan Institute of Oncology-IDIBIGI, 17007 Girona, Spain. ¹¹Centre for Human Genetics, University of Marburg, Marburg, Germany. ¹²Institute of Neuroscience and Medicine (INM-1), Research Center Jülich, Jülich, Germany.

Received: 12 September 2022 Accepted: 21 February 2023

Published online: 05 March 2023

References

- Carr PR, Weigl K, Edelmann D, Jansen L, Chang-Claude J, Brenner H, et al. Estimation of absolute risk of colorectal cancer based on healthy lifestyle, genetic risk, and colonoscopy status in a population-based study. *Gastroenterology*. 2020;159:129–38.
- Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med*. 2000;343:78–85.
- Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer*. 2002;99:260–6.
- Win AK, Jenkins MA, Dowty JG, Antoniou AC, Lee A, Giles GG, et al. Prevalence and penetrance of major genes and polygenes for colorectal cancer. *Cancer Epidemiol Biomarkers Prev*. 2017;26:404–12.
- Biller LH, Syngal S, Yurgelun MB. Recent advances in Lynch syndrome. *Fam Cancer*. 2019;18:211–9.
- Grzymalski JJ, Elhanan G, Morales Rosado JA, Smith E, Schlauch KA, Read R, et al. Population genetic screening efficiently identifies carriers of autosomal dominant diseases. *Nat Med*. 2020;26:1235–9.
- Talseth-Palmer BA. The genetic basis of colonic adenomatous polyposis syndromes. *Hered Cancer Clin Pract*. 2017;15:5.
- Kanth P, Grimmett J, Champine M, Burt R, Samadder NJ. Hereditary colorectal polyposis and cancer syndromes: a primer on diagnosis and management. *Am J Gastroenterol*. 2017;112:1509–25.
- Vogt S, Jones N, Christian D, Engel C, Nielsen M, Kaufmann A, et al. Expanded extracolonic tumor spectrum in MUTYH-associated polyposis. *Gastroenterology*. 2009;137:1976–85.
- Win AK, Dowty JG, Cleary SP, Kim H, Buchanan DD, Young JP, et al. Risk of colorectal cancer for carriers of mutations in MUTYH,

- with and without a family history of cancer. *Gastroenterology*. 2014;146:1208–11.
11. Stoffel EM, Koeppel E, Everett J, Ulintz P, Kiel M, Osborne J, et al. Germline genetic features of young individuals with colorectal cancer. *Gastroenterology*. 2018;154:897–905.
 12. Chubb D, Broderick P, Dobbins SE, Frampton M, Kinnersley B, Penegar S, et al. Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nat Commun*. 2016;7:11883.
 13. Schubert SA, Morreau H, de Miranda NFCC, van Wezel T. The missing heritability of familial colorectal cancer. *Mutagenesis*. 2020;35:221–31.
 14. Yurgelun MB, Kulke MH, Fuchs CS, Allen BA, Uno H, Hornick JL, et al. Cancer susceptibility gene mutations in individuals with colorectal cancer. *J Clin Oncol*. 2017;35:1086–95.
 15. Brenner H, Hoffmeister M, Haug U. Family history and age at initiation of colorectal cancer screening. *Am J Gastroenterol*. 2008;103:2326–31.
 16. Butterworth AS, Higgins JP, Pharoah P. Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis. *Eur J Cancer*. 2006;42:216–27.
 17. McGeoch L, Saunders CL, Griffin SJ, Emery JD, Walter FM, Thompson DJ, et al. Risk prediction models for colorectal cancer incorporating common genetic variants: a systematic review. *Cancer Epidemiol Biomarkers Prev*. 2019;28:1580–93.
 18. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet*. 2019;51:76–87.
 19. Schmit SL, Edlund CK, Schumacher FR, Gong J, Harrison TA, Huyghe JR, et al. Novel common genetic susceptibility loci for colorectal cancer. *J Natl Cancer Inst*. 2019;111:146–57.
 20. Lu Y, Kweon SS, Tanikawa C, Jia WH, Xiang YB, Cai Q, et al. Large-scale genome-wide association study of East Asians identifies loci associated with risk for colorectal cancer. *Gastroenterology*. 2019;156:1455–66.
 21. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun*. 2019;10:2154.
 22. Thomas M, Sakoda LC, Hoffmeister M, Rosenthal EA, Lee JK, van Duijnhoven FJB, et al. Genome-wide modeling of polygenic risk score in colorectal cancer risk. *Am J Hum Genet*. 2020;107:432–44.
 23. Hsu L, Jeon J, Brenner H, Gruber SB, Schoen RE, Berndt SI, et al. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology*. 2015;148:1330–9.
 24. Frampton MJ, Law P, Litchfield K, Morris EJ, Kerr D, Turnbull C, et al. Implications of polygenic risk for personalised colorectal cancer screening. *Ann Oncol*. 2016;27:429–34.
 25. Mur P, Bonifaci N, Díez-Villanueva A, Munté E, Alonso MH, Obón-Santacana M, et al. Non-lynch familial and early-onset colorectal cancer explained by accumulation of low-risk genetic variants. *Cancers*. 2021;13:3857.
 26. Frampton M, Houlston RS. Modeling the prevention of colorectal cancer from the combined impact of host and behavioral risk factors. *Genet Med*. 2017;19:314–21.
 27. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50:1219–24.
 28. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562:203–9.
 29. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc*. 2015;10:1556–66.
 30. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434–43.
 31. Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, Neben CL, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun*. 2020;11:3635.
 32. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42:D980–985.
 33. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7.
 34. Khera AV, Chaffin M, Zekavat SM, Collins RL, Roselli C, Natarajan P, et al. Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation*. 2019;139:1593–602.
 35. Briggs SEW, Law P, East JE, Wordsworth S, Dunlop M, Houlston R, et al. Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: population based cohort study. *BMJ*. 2022;379:e071707.
 36. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018;19:581–90.
 37. Jenkins MA, Makalic E, Dowty JG, Schmidt DF, Dite GS, MacInnis RJ, et al. Quantifying the utility of single nucleotide polymorphisms to guide colorectal cancer screening. *Future Oncol*. 2016;12:503–13.
 38. Jia G, Lu Y, Wen W, Long J, Liu Y, Tao R, et al. Evaluating the utility of polygenic risk scores in identifying high-risk individuals for eight common cancers. *JNCI Cancer Spectr*. 2020;4:21.
 39. Archambault AN, Su YR, Jeon J, Thomas M, Lin Y, Conti DV, et al. Cumulative burden of colorectal cancer-associated genetic variants is more strongly associated with early-onset vs late-onset cancer. *Gastroenterology*. 2020;158:1274–86.
 40. Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE, Willett WC. A prospective study of family history and the risk of colorectal cancer. *N Engl J Med*. 1994;331:1669–74.
 41. Rex DK, Boland CR, Dominitz JA, Giardiello FM, Johnson DA, Kaltenbach T, et al. Colorectal cancer screening: recommendations for physicians and patients from the U.S. Multi-Society Task Force on Colorectal Cancer. *Am J Gastroenterol*. 2017;112:1016–30.
 42. Wijnen JT, Brohet RM, van Eijk R, Jagmohan-Changur S, Middeldorp A, Tops CM, et al. Chromosome 8q23.3 and 11q23.1 variants modify colorectal cancer risk in Lynch syndrome. *Gastroenterology*. 2009;136:131–7.
 43. Talseth-Palmer BA, Wijnen JT, Brenne IS, Jagmohan-Changur S, Barker D, Ashton KA, et al. Combined analysis of three Lynch syndrome cohorts confirms the modifying effects of 8q23.3 and 11q23.1 in MLH1 mutation carriers. *Int J Cancer*. 2013;132:1556–64.
 44. Hassanin E, May P, Aldisi R, Spier I, Forstner AJ, Nöthen MM, et al. Breast and prostate cancer risk: the interplay of polygenic risk, rare pathogenic germline variants, and family history. *Genet Med*. 2022;24:576–85.
 45. Mars N, Widén E, Kerminen S, Meretoja T, Pirinen M, Della Briotta Parolo P, et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat Commun*. 2020;11:6383.
 46. Gao C, Polley EC, Hart SN, Huang H, Hu C, Gnanalivu R, et al. Risk of breast cancer among carriers of pathogenic variants in breast cancer predisposition genes varies by polygenic risk score. *J Clin Oncol*. 2021;39:2564–73.
 47. Jenkins MA, Win AK, Dowty JG, MacInnis RJ, Makalic E, Schmidt DF, et al. Ability of known susceptibility SNPs to predict colorectal cancer risk for persons with and without a family history. *Fam Cancer*. 2019;18:389–97.
 48. Biller LH, Horiguchi M, Uno H, Ukaegbu C, Syngal S, Yurgelun MB. Familial burden and other clinical factors associated with various types of cancer in individuals with lynch syndrome. *Gastroenterology*. 2021;161:143–50.
 49. Kachuri L, Graff RE, Smith-Byrne K, Meyers TJ, Rashkin SR, Ziv E, et al. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat Commun*. 2020;11:6084.
 50. Jenkins MA, Buchanan DD, Lai J, Makalic E, Dite GS, Win AK, et al. Assessment of a polygenic risk score for colorectal cancer to predict risk of lynch syndrome colorectal cancer. *JNCI Cancer Spectr*. 2021;5:22.
 51. Choi J, Jia G, Wen W, Long J, Shu XO, Zheng W. Effects of screenings in reducing colorectal cancer incidence and mortality differ by polygenic risk scores. *Clin Transl Gastroenterol*. 2021;12:e00344.
 52. Stanesby O, Jenkins M. Comparison of the efficiency of colorectal cancer screening programs based on age and genetic risk for reduction of colorectal cancer mortality. *Eur J Hum Genet*. 2017;25:832–8.
 53. Naber SK, Kundu S, Kuntz KM, Dotson WD, Williams MS, Zauber AG, et al. Cost-effectiveness of risk-stratified colorectal cancer screening based on polygenic risk: current status and future potential. *JNCI Cancer Spectr*. 2020;4:86.
 54. Monahan KJ, Bradshaw N, Dolwani S, Desouza B, Dunlop MG, East JE, et al. Guidelines for the management of hereditary colorectal cancer from the British Society of Gastroenterology (BSG)/Association of

- Coloproctology of Great Britain and Ireland (ACPGBI)/United Kingdom Cancer Genetics Group (UKCGG). *Gut*. 2020;69:411–44.
55. Seppälä TT, Dominguez-Valentin M, Sampson JR, Møller P. Prospective observational data informs understanding and future management of Lynch syndrome: insights from the Prospective Lynch Syndrome Database (PLSD). *Fam Cancer*. 2021;20:35–9.
 56. Tyrrell J, Zheng J, Beaumont R, Hinton K, Richardson TG, Wood AR, et al. Genetic predictors of participation in optional components of UK Biobank. *Nat Commun*. 2021;12:886.
 57. Saunders CL, Kilian B, Thompson DJ, McGeoch LJ, Griffin SJ, Antoniou AC, et al. External validation of risk prediction models incorporating common genetic variants for incident colorectal cancer using UK biobank. *Cancer Prev Res*. 2020;13:509–20.
 58. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51:584–91.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



3.3 Publication 3: Assessing the performance of European-derived cardiometabolic polygenic risk scores in South-Asians and their interplay with family history

Contribution: Data analysis, interpretation of results, writing and revision of manuscript

RESEARCH

Open Access



Assessing the performance of European-derived cardiometabolic polygenic risk scores in South-Asians and their interplay with family history

Emadeldin Hassanin^{1,2}, Carlo Maj³, Hannah Klinkhammer^{2,4}, Peter Krawitz², Patrick May^{1†} and Dheeraj Reddy Bobbili^{1,5*†}

Abstract

Background & aims We aimed to assess the performance of European-derived polygenic risk scores (PRSs) for common metabolic diseases such as coronary artery disease (CAD), obesity, and type 2 diabetes (T2D) in the South Asian (SAS) individuals in the UK Biobank. Additionally, we studied the interaction between PRS and family history (FH) in the same population.

Methods To calculate the PRS, we used a previously published model derived from the EUR population and applied it to the individuals of SAS ancestry from the UKB study. Each PRS was adjusted according to an individual's genotype location in the principal components (PC) space to derive an ancestry adjusted PRS (aPRS). We calculated the percentiles based on aPRS and stratified individuals into three aPRS categories: low, intermediate, and high. Considering the intermediate-aPRS percentile as a reference, we compared the low and high aPRS categories and generated the odds ratio (OR) estimates. Further, we measured the combined role of aPRS and first-degree family history (FH) in the SAS population.

Results The risk of developing severe obesity for SAS individuals was almost twofold higher for individuals with high aPRS than for those with intermediate aPRS, with an OR of 1.95 (95% CI = 1.71–2.23, $P < 0.01$). At the same time, the risk of severe obesity was lower in the low-aPRS group (OR = 0.60, CI = 0.53–0.67, $P < 0.01$). Results in the same direction were found in the EUR data, where the low-PRS group had an OR of 0.53 (95% CI = 0.51–0.56, $P < 0.01$) and the high-PRS group had an OR of 2.06 (95% CI = 2.00–2.12, $P < 0.01$). We observed similar results for CAD and T2D. Further, we show that SAS individuals with a familial history of CAD and T2D with high-aPRS are associated with a higher risk of these diseases, implying a greater genetic predisposition.

Conclusion Our findings suggest that CAD, obesity, and T2D GWAS summary statistics generated predominantly from the EUR population can be potentially used to derive aPRS in SAS individuals for risk stratification. With future

[†]Patrick May and Dheeraj Reddy Bobbili contributed equally.

*Correspondence:
Dheeraj Reddy Bobbili
dheeraj.bobbili@uni.lu

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

GWAS recruiting more SAS participants and tailoring the PRSs towards SAS ancestry, the predictive power of PRS is likely to improve further.

Keywords Type 2 diabetes, Family History, South Asians, Polygenic risk, Coronary artery disease

Background

Several genome-wide association studies (GWAS) for more than 5000 traits in GWAS Catalog [1] have been conducted to date, and very few of the GWASs have had significant success translating into the clinical setting [2]. Hence, it is a significant milestone to translate GWAS findings to clinical settings, particularly for traits with high heritability. One of the drawbacks of the GWAS findings is that the identified genome-wide significant SNPs do not have such a large effect size in most cases. However, a current approach of combining those SNPs to a single score known as a polygenic risk score (PRS) has become popular to enhance the accuracy of predicting individuals at risk and has thus shifted the focus of the genetic community towards the use of GWAS findings again [3]. PRS can be a precious tool for risk stratification, particularly in identifying groups of people with extremely high or low genetic risk of developing a specific disease or trait. Moreover, based on our recent work and others, it has become clear that for certain traits, high PRS, along with rare disease-causing variants, can further increase the individuals' risk of developing a disease compared to carriers without a high PRS [4–7].

Identifying the risk SNPs using GWAS requires a considerable sample size as even most disease-related SNPs have relatively small effect sizes. So far, most of the larger GWASs have been mainly conducted in individuals with European (EUR) ancestries. One of PRS limitations is that it may not be transferable between different ancestries [8]. Due to both potential gene–environment interactions and population structure the application of EUR GWAS derived PRS can be problematic in non-EUR populations as it often results in shifted PRS distribution [8]. This lack of portability of PRS is due to differences in linkage disequilibrium (LD), risk variants, effect sizes, and allele frequencies. Further, methods to genotype or impute the missing SNPs initially developed with samples of EUR ancestry can increase those differences [9]. The critical demand to advance polygenic prediction in non-European populations is not being met, as South Asian (SAS) groups, which form the largest ancestry group encompassing 23% of the world's population [10], remain significantly underrepresented in existing GWAS studies. This underscores the imperative to substantially increase their participation in genetic research [11].

Despite ongoing efforts to increase global genetics research diversity, it will take still some time to attain sufficient GWAS sample sizes to identify population-specific risk SNPs. As mentioned earlier, PRS is a potent tool to

identify the sub-populations at risk. However, this inability to use it across populations with different ancestries is an important research topic. Several studies were being performed to study the portability of EUR-derived PRSs into other ancestries and an SAS specific PRS has been developed for CAD using previously published GWAS statistics [10]. However, the majority of them had limited success [12–14]. The PRS derived from EUR performed poorly in African population [15] and similar results were observed in a Latino/Hispanic population for some traits [16]. While EUR-derived PRSs showed similar results for quantitative traits like blood count and anthropometric features, it performed poorly for blood pressure traits [17]. Others have shown a connection between PRS and genetic ancestry [12, 18]. In other words, the studies show that applying PRSs derived from the EUR population directly on other ancestries might not be ideal. However, few studies used an approach to developing an ancestry-adjusted PRS (aPRS) that is mainly derived from EUR and can be transferred to other ethnicities [19]. For example, a study showed a compromised solution where they found a minimal decrease in the prediction power of the PRS in SAS compared to EUR [20].

Recently, it has been shown that in breast cancer, the PRS derived from EURs with an ancestry correction performed well in the SAS population [14]. However, it is still unclear to what extent populations of EUR and SAS ancestry share the same genetic underpinnings of such cardiometabolic/lifestyle traits, and such an assessment is still missing. It is of utmost importance to perform this assessment because compared to other ethnicities, SAS individuals have an increased susceptibility to coronary artery disease (CAD), obesity, and type 2 diabetes (T2D) [21]. The interplay between PRS and family history (FH) in predicting the risk of various diseases has been a topic of interest in recent years [5, 22–24]. Although previous studies have examined the independent effects of FH and PRS, there is a lack of systematic research on the relative contributions and overlap of these factors across different types of familial risk in SAS.

Here, we systematically assessed the portability of the aPRS derived from EUR ancestry for obesity, CAD, and T2D to the SAS population and the interplay of FH and PRS in the same individuals. Hence, we used a published list of SNPs derived from the PGS catalog [25], then generated the aPRS and applied it to the EUR and SAS samples from the UK Biobank (UKB).

Methods

Data source

The UKB is a prospective study that collects data over a long period and recruits volunteers aged between 40 and 69, mostly from Scotland, Wales, and England, totaling over 500,000 individuals. All participants have provided written consent and collected data is available for research purposes. The UK Biobank Axiom Array was used to generate genotyping data, which included around 850,000 variants and the imputation of over 90 million variants [26].

Study cohort

CAD and T2D diagnoses were based on self-reported illness codes and international Classification of Diseases (ICD)-10 and ICD-9 diagnosis codes, and Office of Population Censuses and Surveys (OPCS-4) procedure codes [3]. CAD was defined using ICD-10 codes (I21.X, I22.X, I23.X, I24.1, or I25.2), ICD-9 codes (410.X, 411.0, 412.X, or 429.79), OPCS-4 codes (K40.[1–4], K41.[1–4], K45.[1–5], K49.[1–2], K49.[8–9], K50.2, K75.[1–75.4], or K75.[8–0.9]), and self-reported illness codes 1075. T2D was defined using ICD-10 code E11.X, and self-reported illness codes 1223. Diagnosis of obesity was based on body mass index (BMI), with individuals having a BMI > 25 considered obese.

We then estimated genetic ancestries (EUR, and SAS) by projecting the samples in the 1000 genome project (1KGP) principal component (PC) spaces, while considering 1KGP superpopulations as a reference. The UKB conducted quality control for the genetic data, and the UKB processed files were used in downstream analysis. We analyzed individuals of EUR and SAS ancestry, and samples with discordant genotypic versus reported sex, sex chromosome aneuploidy, and high heterozygosity or missing genotype rates were considered as outliers (coded as “YES” in the fields 22,001, 22,019, and 22,027 respectively) and excluded from further analysis. We included only individuals who are unrelated up to the second degree, and from each pair of related individuals, one member was randomly retained (kinship coefficient > 0.0884, according to the UKB).

Polygenic risk score analysis

PRSs were calculated using panels of SNPs identified in the previous studies [3, 27] and the effect sizes were downloaded from PGS catalog [21] using the ids PGS000027, PGS000013, PGS000014 for BMI, CAD and T2D respectively. PRSice-2 was used to generate the PRS, which account automatically for allele-flipping and removing ambiguous SNPs [28]. Strand-ambiguous SNPs are the ones with A/T or C/G alleles. Since many GWAS studies do not report the strand assignments, it is a standard practice in PRS calculations to exclude ambiguous

SNPs. Since we already obtained the list of SNPs for the PRS calculation, we utilized the ‘*–no clumping*’ and ‘*–no regress*’ parameters along with the other default parameters, to bypass the time-consuming steps of regression and clumping. PRS values were standardized using the mean and standard deviation for the whole data.

Adjustment of PRS

Based on an previously applied approach [5, 19] to reduce the variation in the PRS distribution due to genetic ancestry, we calculated an adjusted PRS (aPRS). A linear regression model was fitted using the PRS as the outcome variable and the first four PC derived from UKB as covariates (PRS ~ PC1 + PC2 + PC3 + PC4). A predicted PRS was calculated based on this model. Finally, the aPRS was calculated by subtracting the predicted PRS from the raw PRS and standardized using the mean and standard deviation.

Statistical analysis

To investigate the association of aPRS and disease risk, we used logistic regressions with the occurrence of the disease as an outcome, i.e., separate logistic regressions for CAD, T2D, and obesity, respectively. All analyses were done for SAS and EUR populations separately.

First, we used aPRS as a continuous variable and adjusted the model for age, sex, and the first four PCs corresponding to the model

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_{aPRS} aPRS + \beta_{sex} sex + \beta_{age} age + \sum_{k=1}^4 \beta_{PC_k} PC_k$$

with $Y = 1$ corresponding to the occurrence of the disease (CAD, T2D or obesity). Adjusted odds ratios (ORs) were calculated as $OR = \exp(\beta_{aPRS})$.

Second, we categorized the aPRS into three groups: low aPRS, intermediate aPRS, and high aPRS. We used the percentiles of the aPRS distribution in the SAS and EUR populations, respectively. SAS individuals were assigned to the “low” aPRS group if their aPRS fell below the 20th percentile (“< 20%”) of the aPRS distribution in the SAS population and to the “high” aPRS group if their aPRS fell above the 80th percentile (“> 80%”) of the aPRS distribution in the SAS population. The remaining SAS individuals were assigned to the “intermediate” aPRS group (“20%–80%”). The same was done for EUR individuals based on the aPRS distribution in the EUR population.

Then we replaced the continuous aPRS variable in the logistic regression by the aPRS group using the “intermediate” aPRS group as the reference category, i.e., we used the model

$$\begin{aligned} \text{logit}(P(Y = 1)) = & \beta_0 + \beta_{aPRS_{low}} aPRS_{low} \\ & + \beta_{aPRS_{high}} aPRS_{high} + \beta_{sex} sex \\ & + \beta_{age} age + \sum_{k=1}^4 \beta_{PC_k} PC_k \end{aligned}$$

with $aPRS_{low} = 1$, if the individual is in the low aPRS group and 0 otherwise (analogous for $aPRS_{high}$). Adjusted ORs for disease occurrence when being in the low or high aPRS group compared to the intermediate aPRS group were calculated as $OR = \exp(\beta_{aPRS_{low}})$ and $OR = \exp(\beta_{aPRS_{high}})$ respectively.

Finally, to determine the combined effect of aPRS and family history (FH), we reclassified the three aPRS groups into six groups based on FH status. FH was defined as positive (and encoded as $FH = pos$) or negative (encoded as $FH = neg$) whether the individual has FH of the corresponding disease in parents or siblings. For example, individuals with high aPRS and positive family history are encoded as $aPRS_{high} FH_{pos}$. We then fitted the logistic regression model

$$\begin{aligned} \text{logit}(P(Y = 1)) = & \beta_0 + \beta_{aPRS_{low} FH_{pos}} aPRS_{low} FH_{pos} \\ & + \beta_{aPRS_{low} FH_{neg}} aPRS_{low} FH_{neg} \\ & + \beta_{aPRS_{int} FH_{pos}} aPRS_{int} FH_{pos} \\ & + \beta_{aPRS_{high} FH_{neg}} aPRS_{high} FH_{neg} \\ & + \beta_{aPRS_{high} FH_{pos}} aPRS_{high} FH_{pos} \\ & + \beta_{sex} sex + \beta_{age} age + \sum_{k=1}^4 \beta_{PC_k} PC_k. \end{aligned}$$

The reference category is then given by individuals with intermediate aPRS and without positive FH. The adjusted OR for the occurrence of disease of individuals with, e.g., high aPRS and positive FH compared to the reference category is then estimated by $OR = \exp(\beta_{aPRS_{high} FH_{pos}})$.

Model performance

For assessing the performance of the different models, the area under the curve (AUC) was used. The R package pROC was used to compute the AUC with 95% confidence intervals (CIs), and AUC. We randomly divided the data into (75%) training and (25%) testing datasets. Logistic regression models were fitted on the training data set, and model prediction and AUC calculations were made using the testing data set by applying the corresponding models. Additionally, we measured the area under Precision-Recall (PR) curve (AUPRC) using the R package PRROC to address the challenge of imbalanced datasets. Since case-controls ratios for T2D, and CAD were substantially higher in the SAS than EUR samples, for additional validation of our models we performed down sampling for SAS population to achieve the same

case-control ratios. While for obesity, case-control ratio was roughly the same between both populations.

Survival analysis

To calculate the cumulative lifetime risk based on aPRS strata and FH status, a Cox proportional hazard model was used. Again, separate models were fitted for each phenotype (CAD, T2D, and obesity) respectively. The occurrence of the disease was considered as the event variable. At the same time, age served as the time scale, i.e., the age at diagnosis was considered as event time for observed cases and the age at the most recent visit for censored control. Adjusted survival curves were produced considering the aPRS group, age, sex, FH, and the first four ancestry PCs. We used the Schoenfeld individual test to test the proportional hazard assumption for each variable. R packages *survival* and *survminer* were used to perform Cox proportional hazard models and test the proportional hazard assumption, and R 4.2.2 was used for all statistical calculations.

Results

UK biobank dataset description

We identified a total of 24,156 CAD cases among individuals of EUR ancestry and 822 SAS cases, with a mean age of 61.51 and 58.71 years at recruitment, respectively. The remaining individuals were considered controls. For T2D, we identified 25,526 cases among EUR individuals and 1,718 cases among SAS individuals, with a mean age of 60.39 and 57.42 years, respectively. For obesity (BMI > 25), we identified 301,385 EUR and 5,690 SAS cases, with a mean age of 55.73 and 53.80 years, in EUR and SAS, respectively (Table 1).

In the SAS population, CAD cases were more common in individuals a positive FH of CAD than individuals without FH of CAD with OR 1.98 [1.70–2.31], $P < 0.01$. Moreover, T2D was diagnosed significantly more frequently in individuals with a positive FH of T2D than in individuals without FH of T2D (OR = 2.09 [1.86–2.34], $P < 0.01$).

Ancestry correction and PRS distribution within the UKBB cohort

When studying individuals of a particular ancestry, it is crucial to apply ancestry correction using principal components (PCs) derived from the reference population. Figure 1 illustrates the effect of this step; while the PRS distributions are shifted horizontally for EUR and SAS populations, the ancestry correction ensures zero-centered aPRS distributions for each population. However, when using only PRS without ancestry correction, we observed a striking difference in the number of individuals assigned to high PRS (where high PRS was defined as an individual belonging to a PRS percentile > 80%).

Table 1 Characteristics of the participants by CAD, T2D, and Obesity diagnosis. Coronary artery disease (CAD), type 2 diabetes (T2D), European (EUR), South Asian (SAS)

Ethnicity	CAD						T2D						Obesity					
	EUR		SAS		EUR		SAS		EUR		SAS		EUR		SAS			
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls		
Participants, N	24,156	428,610	822	7842	25,526	427,240	1718	6946	301,385	149,889	5690	2773	301,385	149,889	5690	2773		
Male, N (%)	18,514 (76.3)	188,835 (44.06)	690 (83.94)	3959 (50)	15,756 (61.3)	191,593 (44.84)	1055 (61.41)	3594 (51)	154,908 (51.39)	51,703 (15.1)	2993 (52.60)	1493 (53.84)	154,908 (51.39)	51,703 (15.1)	2993 (52.60)	1493 (53.84)		
Female, N (%)	5642 (23.36)	239,775 (55.94)	132 (16.06)	3883 (49)	9770 (38)	235,647 (55.16)	663 (38.59)	3352 (48)	146,477 (48.61)	98,186 (28.7)	2697 (47.40)	1280 (46.16)	146,477 (48.61)	98,186 (28.7)	2697 (47.40)	1280 (46.16)		
Age, mean (SD)	61.51 (6.19)	56.53 (8.03)	58.71 (7.69)	53.05 (8.35)	60.39 (6.77)	56.58 (8.04)	57.42 (7.79)	52.64 (8.34)	55.73 (7.64)	56.81 (8.02)	53.8 (8.36)	53.1 (8.65)	55.73 (7.64)	56.81 (8.02)	53.8 (8.36)	53.1 (8.65)		
Family history of CAD, N (%)	14,759 (61.2)	188,340 (43.94)	457 (55.6)	3193 (40)	13,131 (50.2)	189,968 (44.46)	754 (43.89)	2896 (41)	138,185 (45.84)	64,259 (20.1)	2411 (42.37)	1159 (41.79)	138,185 (45.84)	64,259 (20.1)	2411 (42.37)	1159 (41.79)		
Family history of T2D, N (%)	5290 (21.9)	90,086 (21.02)	431 (52.43)	4144 (52)	10,120 (39.6)	85,256 (21.0)	1113 (64.78)	3462 (49)	68,794 (20.1)	26,267 (7.7)	3059 (53.76)	1430 (51.56)	68,794 (20.1)	26,267 (7.7)	3059 (53.76)	1430 (51.56)		

Specifically, there were significant variations between ethnic groups (EUR and SAS). In cases where matched reference controls are available, ancestry correction might not be necessary. However, due to the underrepresentation of SAS populations in current genetic studies, it is crucial to explore alternative approaches when ancestry-matched reference controls are not accessible. This will ensure more accurate and applicable results for diverse populations. For example, 18.5% of EUR samples (83,955/452,766) had a high PRS, while almost all SAS samples (96.2%, 8,331/8,664) showed a high PRS. However, applying aPRS reduced this variation. For instance, 20% of EUR samples (90,627/452,766) and 19.2% of SAS samples (1,659/8,664) had a high aPRS, leading to a more comparable distribution of PRS across ethnic groups. Similar results have been observed for CAD and obesity as well (Table 2). Our findings are in line with a previous study where they showed that ancestry correction is crucial to place an individual in the correct aPRS percentile for disease risk prediction [20].

Performance of aPRS on SAS individuals and association with disease development

Our analysis revealed that the models incorporating both adjusted polygenic risk scores (aPRS) and covariates have improved performance compared to the models based solely on covariates. This was evident from the higher AUC values for all three conditions - obesity, coronary artery disease (CAD), and type 2 diabetes (T2D) - when aPRS was included in covariates models. Specifically, for obesity, the AUC increased from 0.56 (95% CI, 0.55–0.57) to 0.63 (95% CI, 0.62–0.86); for CAD, it rose from 0.76 (95% CI, 0.75–0.78) to 0.79 (95% CI, 0.77–0.8); and for T2D, it increased from 0.67 (95% CI, 0.66–0.68) to 0.69 (95% CI, 0.68–0.7) (Fig. 2). These improvements in AUC values suggest that incorporating aPRS into the models enhances their ability to discriminate between cases and controls for obesity, CAD, and T2D. Following the downsampling process outlined in the methods section, we did not identify any substantial differences in the performance of the model Supplementary Fig. 1.

Additionally, we observed improvements in the Area Under the Precision-Recall Curve (AUPRC) values for all three conditions when aPRS was incorporated into the models. The detailed AUPRC values can be found in Supplementary Fig. 2, which highlights the enhanced precision-recall balance achieved by including aPRS in the models. This further supports the conclusion that aPRS is a valuable addition to the models for predicting the risk of obesity, CAD, and T2D. AUROC and AUPRC values are provided in Supplementary Fig. 2. The models performance in EUR and SAS showed similar trends for AUROC Supplementary Fig. 3 and AUPRC Supplementary Fig. 4.

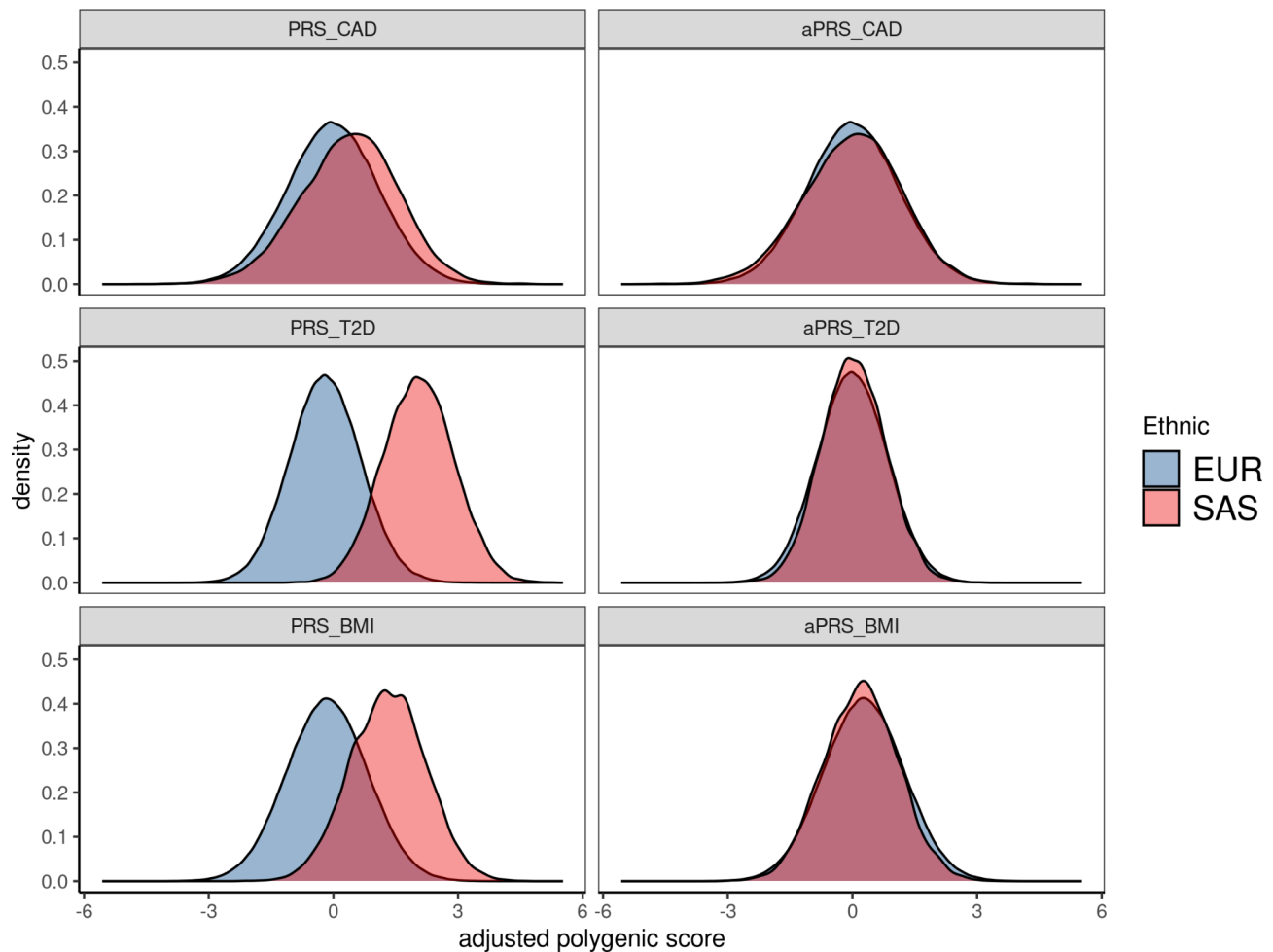


Fig. 1 The distribution of PRSs before and after ancestry corrections across the various diseases. European (EUR), South Asian (SAS), coronary artery disease (CAD), type 2 diabetes (T2D), and adjusted polygenic risk scores (aPRS).

Table 2 Comparison of the distribution of high (a) PRS (defined as PRS percentile > 80%). Coronary artery disease (CAD), type 2 diabetes (T2D), European (EUR), South Asian (SAS), Adjusted (a) polygenic risk scores (PRS)

	Ancestry correction	EUR samples with High PRS	SAS samples with High PRS
T2D	PRS	83,955 (18.5%)	8331 (96.2%)
	adjusted PRS (aPRS)	90,627 (20%)	1659 (19.2%)
CAD	PRS	89,438 (19.8%)	2,848 (32.9%)
	adjusted PRS (aPRS)	90,440 (20%)	1,846 (21.3%)
Obesity	PRS	85,853 (19%)	6433 (74.3%)
	adjusted PRS (aPRS)	90,794 (20.1%)	1492 (17.2%)

Disease association with aPRS categorization in South Asians

Our investigation into the performance of aPRS on SAS individuals revealed an increasing in the risk of

developing coronary artery disease (CAD) based on aPRS categorization. Individuals with a low aPRS demonstrated significantly reduced odds of developing CAD, with an odds ratio (OR) of 0.56 (95% CI: 0.45–0.7), indicating a lower risk than the reference group. Conversely, those with a high aPRS exhibited an elevated CAD risk, with an OR of 1.72 (95% CI: 1.44–2.05).

Similarly, in the SAS population, the association between aPRS categorization and obesity risk showed similar results. Individuals in the high aPRS group had an OR of 1.95 (95% CI=1.71–2.23) compared to those in the intermediate aPRS group. Regarding type 2 diabetes (T2D), the high aPRS group in the SAS population had an OR of 1.55 (95% CI, 1.36–1.77) (Fig. 3). While comparing with the EUR individuals a similar trend has been observed Supplementary Fig. 5.

Association of CAD and T2D with family history and aPRS

Individuals with a positive FH and high aPRS showed a higher risk of developing CAD than those with no FH

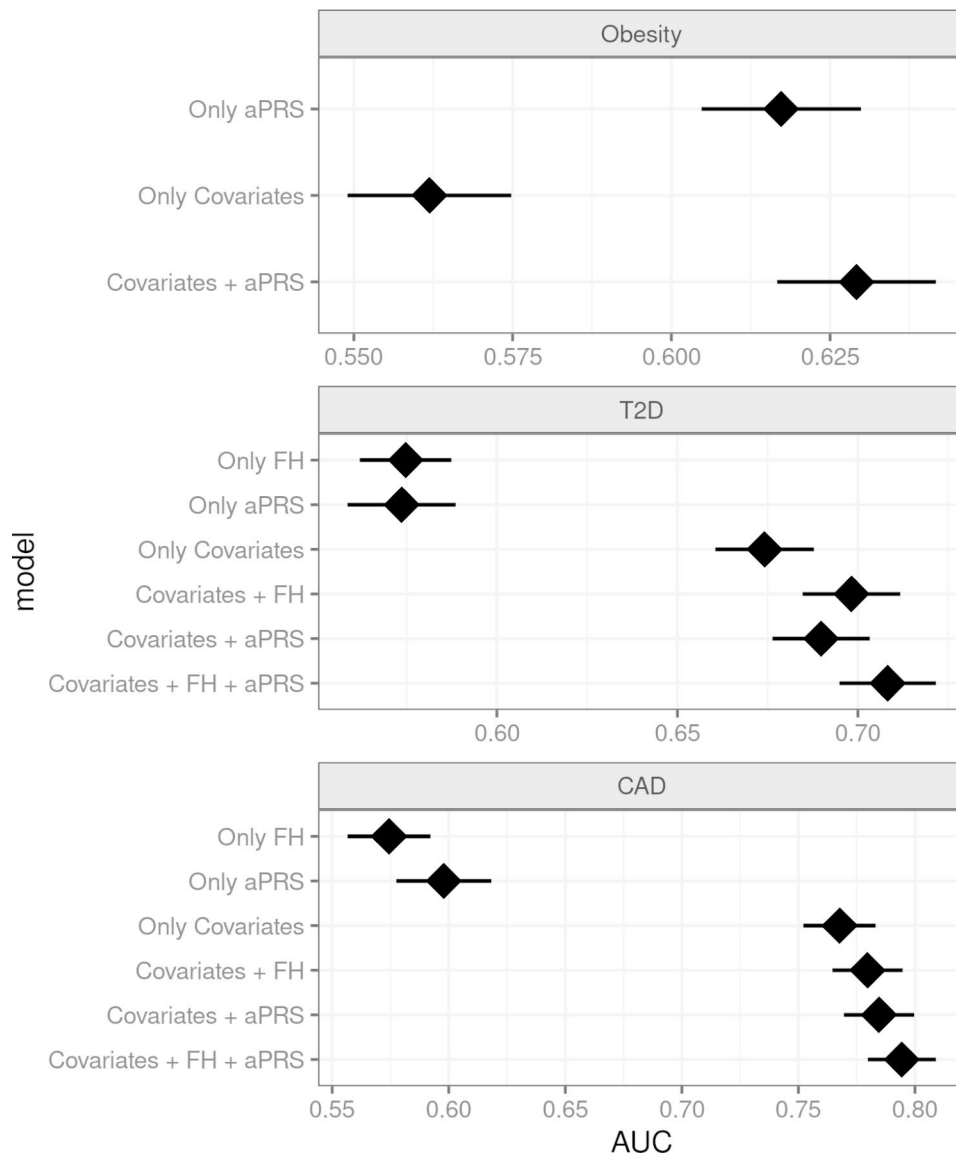


Fig. 2 Comparison of different models and their corresponding AUCs among South-Asian (SAS) population. Ancestry adjusted PRS (aPRS), First degree family history (FH) and covariates (age, sex, first four principal components). European (EUR), coronary artery disease (CAD), type 2 diabetes (T2D).

and intermediate aPRS (Fig. 4). In SAS, those with both positive FH and high aPRS had a more than three-fold increased chance of developing CAD compared to those with intermediate aPRS and no FH, while individuals with a low aPRS and no FH showed a reduced chance of developing CAD with an OR of 0.63 (95%,0.48–0.91). No significant interaction was observed between FH status and PRS $p=0.11$, respectively) (Fig. 4). Notably, in both SAS and EUR, individuals with negative FH and high aPRS had comparable risks of developing CAD as those with positive FH and intermediate aPRS (2-fold risk) Supplementary Fig. 6. The same trend was also shown in T2D.

Cox-proportional hazard analysis

For the cox-proportional hazard, the Schoenfeld tests conducted on each covariate and the global test do not yield statistically significant results. Consequently, we can reasonably conclude that the assumption of proportional hazards is not violated Supplementary Table 1.

The cumulative CAD incidence among SAS with positive FH increased from 46% with low aPRS to 75% with high aPRS by age 70 (Fig. 5). Notably, SAS individuals with an intermediate aPRS and a positive FH had a cumulative CAD incidence by age 70 (65%) comparable to those with a high aPRS and a negative FH (63%). The cumulative incidence of T2D among SAS individuals ranges from 58% with a negative FH and low aPRS to 95% with a positive FH and high aPRS (Fig. 5). The cumulative

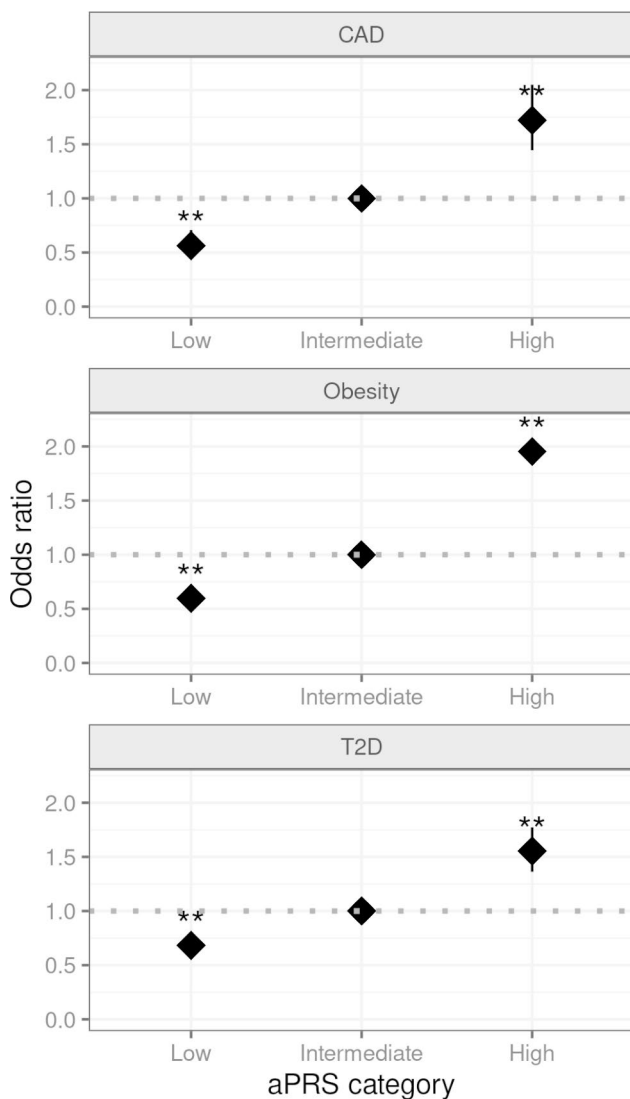


Fig. 3 Odds ratio for CAD, and T2D based on the categorization of based on the adjusted polygenic risk scores (aPRS) percentile in the South Asian (SAS) population of the UK Biobank. Coronary artery disease (CAD), type 2 diabetes (T2D). If a p-value is less than 0.01, it is flagged with two stars (**)

incidence of T2D among individuals with high aPRS of SAS ancestry (95%) was higher than EUR individuals (70%) in the corresponding aPRS groups Supplementary Fig. 7.

Discussion

Extending the previous studies, we aimed to assess the performance of EUR-derived PRSs in the SAS population and explore the relationship between PRS and FH in contributing to the burden of CAD, T2D, and obesity. The results of this study, utilizing UK Biobank data, suggest that an aPRS derived from a large-scale GWAS of cardiometabolic diseases in individuals of European (EUR) ancestry could potentially identify those with an elevated risk of disease predisposition in the South Asian (SAS)

population, albeit with a reduced performance observed in the EUR ancestry group. Additionally, the aPRS may identify SAS individuals with increased risk for T2D and CAD independent of their FH. Among high aPRS individuals with positive FH, we noticed an increased cumulative incidence in individuals of SAS ancestry compared to EUR individuals stratified by PRS (Fig. 5).

It has been shown that the UKB is a valuable resource for evaluating the utility of PRS, as it provides both phenotypic and genotypic data [29]. While most UKB participants have EUR ancestry, the dataset involves more than 20,000 participants of self-reported non-EUR.

However, a major challenge with using PRS in clinical settings is that the distribution of genetic variants can vary widely among different ethnic populations [8]. This can result in inaccurate disease risk predictions and hinder the validation of PRS in diverse populations (see Fig. 1). The observed dissimilarity between the distributions for EURs and SASs highlights the need to adjust for the correct ancestral background to accurately assign an individual to their respective percentile within the reference distribution.

We have used population structure adjustment [20] to address this issue, accounting for the genetic differences between different populations when calculating PRS. By adjusting for population structure, we minimized the impact of genetic variability on the accuracy of PRS predictions and facilitate the validation of PRS in diverse populations.

The generalizability of the study's findings is subject to limitations stemming from several factors. The study participants were recruited exclusively within the UK, including individuals of EUR and SAS ancestry. Thus, healthcare access and non-genetic risk factors may be more comparable among these ethnic groups as they would be expected using two cohorts recruited in EUR and SAS separately. Nevertheless, it is important to acknowledge that socioeconomic determinants, lifestyle choices, and health disparities may differ across various ethnic groups, even living in the same region. Although certain risk variants are likely specific to certain populations, the findings indicating the similar performance of the PRS across ancestry groups suggest that non-EUR groups, including SAS, may share some of the identified risk variants found in EUR-based GWAS for cardiometabolic disorders.

The findings of our study reveal that a higher PRS was associated with an increase in obesity, T2D, and CAD cases among individuals of SAS ancestry. However, the performance of the EUR-based PRSs was less effective in the African (AFR) population, suggesting the existence of ancestry-specific differences [30]. Hence, PRSs should be evaluated carefully by ancestry groups to assess their transferability across ancestries and diseases. Whenever

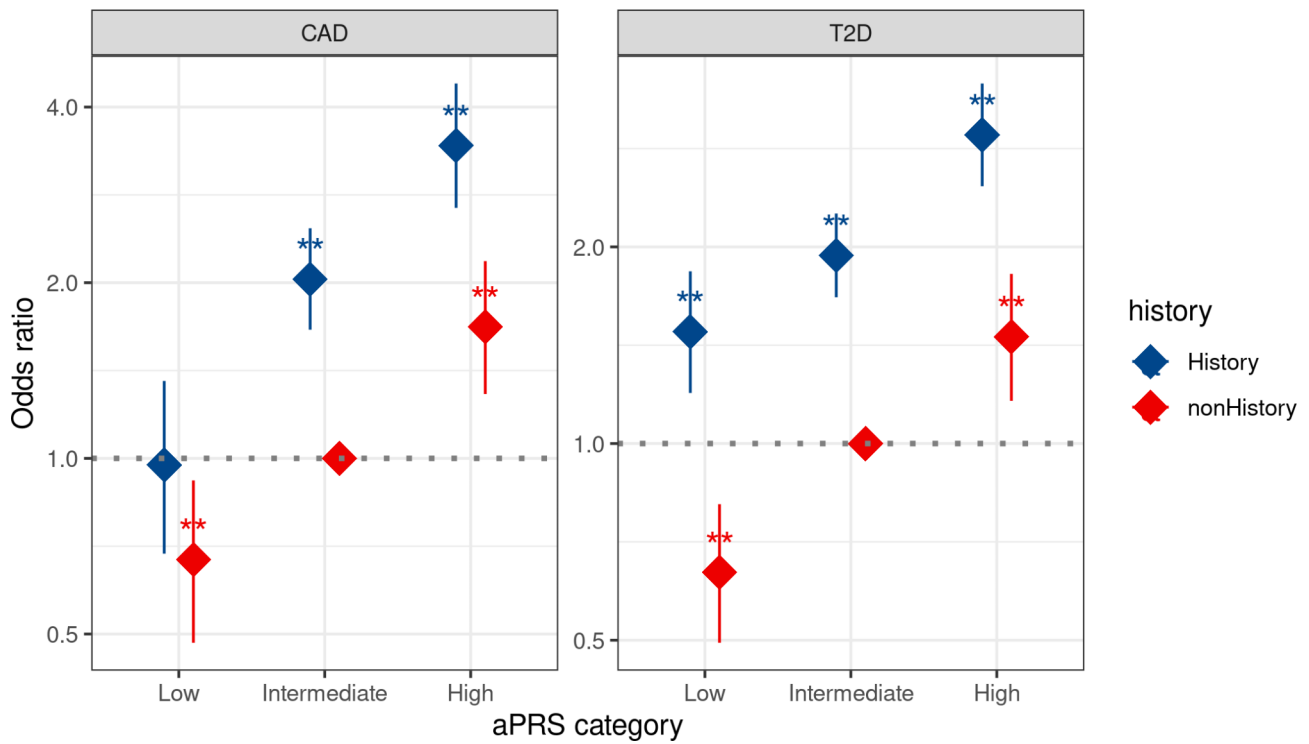


Fig. 4 Odds ratio for CAD, and T2D based on the categorization of based on the adjusted polygenic risk scores (aPRS) percentile and family history (FH) status in the South Asian (SAS) and European (EUR) population of the UK Biobank. Coronary artery disease (CAD), type 2 diabetes (T2D), and adjusted polygenic risk scores (aPRS). If a p-value is less than 0.01, it is flagged with two stars (**)

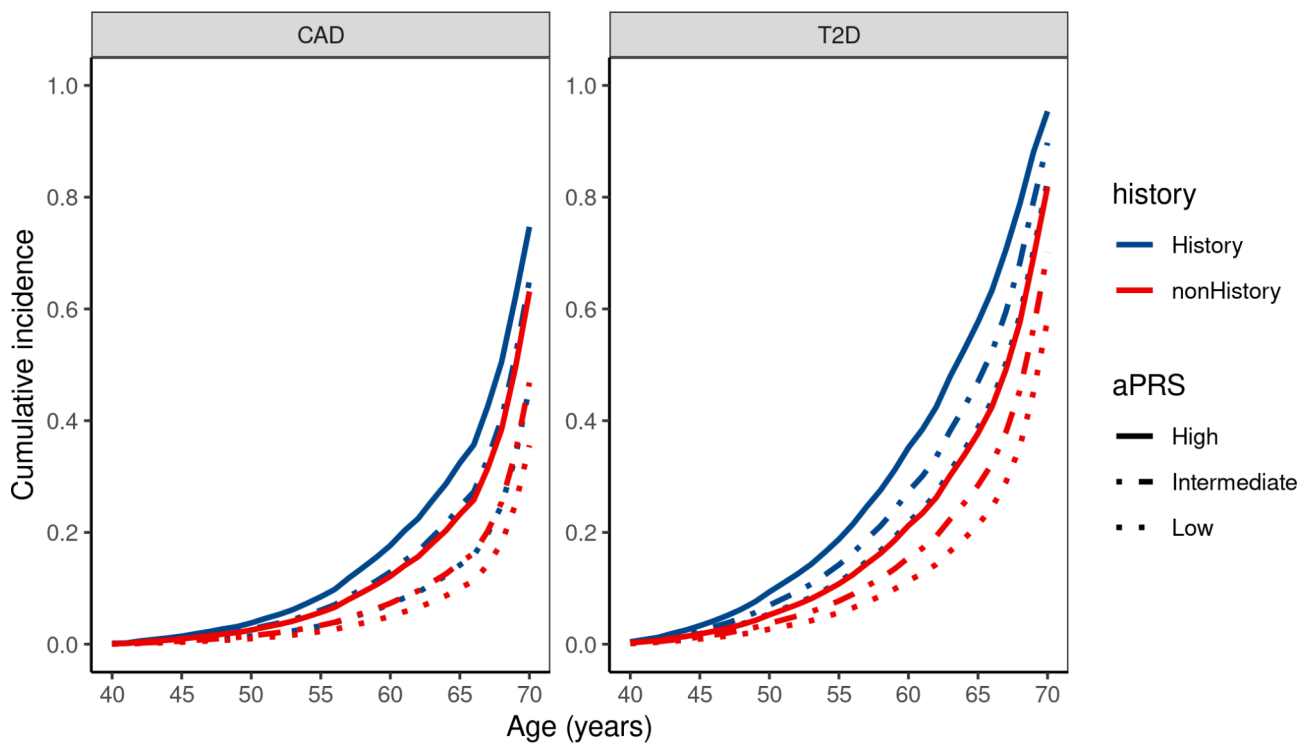


Fig. 5 Cumulative incidence of CAD, T2D, and obesity based on the categorization of based on the adjusted polygenic risk scores (aPRS) percentile and family history (FH) status in the South Asian (SAS) population of the UK Biobank. Coronary artery disease (CAD), type 2 diabetes (T2D).

possible, PRS should be constructed based on GWAS based on the same ancestry group [31].

PRS derived from EUR GWAS may not be optimal for all diseases in non-EUR populations, but they can still offer some value in risk assessments for specific conditions [32]. Postponing implementation until ancestry-specific GWAS or multi-ancestry meta-analyses become available could unintentionally widen health disparities across various populations. In the meantime, while larger non-European cohorts are being established, our study illustrates that employing an adjusted PRS based on a EUR GWAS population can provide a limited level of risk categorization for metabolic traits in SAS individuals. However, additional validation is required to ascertain its efficacy.

The increasing availability of data from larger and more diverse populations, coupled with technological advancements, has spurred interest in the clinical adoption of PRS. Recent research has demonstrated that combining clinical risk scores with PRS can help identify more people at risk of developing T2D, especially in SAS populations. Our study provides a potential model for laboratories and health systems seeking to utilize a EUR-derived PRS in SAS populations. Additionally, our study contributes to the literature that supports using PRS and FH as complementary measures in assessing inherited disease susceptibility for T2D and CAD [5].

One of the key findings of our study is the potential improvement in risk prediction when combining family history with PRS [33]. Several theoretical bases support this notion. Family history might reflect the presence of rare genetic variants that are not included in PRS as they are typically constructed from common genetic variants. Additionally, family members often share similar environments and lifestyles, which can contribute to disease risk and may be captured by family history. This shared environment can also influence gene-environment interactions, another potential risk factor for disease. Furthermore, the disease penetrance, or the likelihood that an individual carrying a particular genetic variant will manifest the disease, can also be impacted by family history. Integrating PRS and family history can offer a more holistic estimate of disease risk, encompassing additional genetic and environmental factors. However, the degree to which this combination improves risk estimation depends on the disease and populations under study.

Conclusion

Taken together, our study provides preliminary evidence that EUR-derived PRSs might be useful to identify individuals at high risk of T2D, obesity, and CAD in the SAS populations. With future GWAS recruiting more SAS participants and tailoring the PRSs towards SAS ancestry, the predictive power of PRS is likely to improve

further. Further, we explored the importance of considering both polygenic risk and family history in assessing disease risk in clinical practice. Such an integration could potentially improve risk prediction and provide personalized prevention and management strategies for the common non-communicable diseases. Further research is needed to assess the clinical utility and cost-effectiveness of implementing these measures in diverse populations.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12920-023-01598-5>.

Supplementary Material 1

Acknowledgements

UK Biobank analyses were conducted via application 52446 using a protocol approved by the Partners HealthCare Institutional Review Board.

Authors' contributions

EH, HK and DRB performed the statistical analysis and the bioinformatics. EH and DRB conceived and designed the study. EH, HK, PM, and DRB drafted the initial manuscript. EH, HK, CM, PK, PM, and DRB performed the critical expert revision. PK, CM, PM, and DRB supervised the study. All authors read and approved the final manuscript.

Funding

PM received funding from the Luxembourg National Research Fund (FNR) INTER grant 'ProtectMove' (INTER/DFG/19/14429377). PM, EH and DRB were supported by FNR (Research Unit FOR-2715, FNR grant INTER/DFG/21/16394868 MechEPI2).

Data Availability

Genome-wide genotyping data, exome-sequencing data, and phenotypic data from the UK Biobank are available upon successful project application (<http://www.ukbiobank.ac.uk/about-biobank-uk/>). Restrictions apply to the availability of these data, which were used under license for the current study (Project ID: 52,446). Please contact the corresponding author for any data related queries. Data are however available from UK Biobank (see <https://www.ukbiobank.ac.uk/enable-your-research> for the application procedure).

Declarations

Ethics approval and consent to participate

This study made use of anonymized data from UK Biobank, which consists of approximately 500,000 volunteers between the ages of 40 and 69 who were gathered throughout Great Britain between 2006 and 2010. Written informed consent was given by all individuals. It was inappropriate to get parental or guardian approval because no one under the age of 16 was recruited. The research ethics committee for the UK Biobank gave its approval to the protocol and the permission. Our investigation was carried out in accordance with approved UK Biobank data application number 52446. We hereby confirm that, all methods were carried out in accordance with relevant guidelines and regulations.

Authors' contributions

EH, HK and DRB performed the statistical analysis and the bioinformatics. EH and DRB conceived and designed the study. EH, HK, PM, and DRB drafted the initial manuscript. EH, HK, CM, PK, PM, and DRB performed the critical expert revision. CM, PK, PM, and DRB supervised the study. All authors read and approved the final manuscript.

Consent for publication

Not Applicable.

Competing interests

DRB is the founder and CEO of Wellytics Technologies Pvt Ltd. No potential conflicts (financial, professional, or personal) for the other authors relevant to the manuscript.

Author details

¹Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 avenue du Swing, Belvaux L-4367, Luxembourg

²Institute for Genomic Statistics and Bioinformatics, University of Bonn, Bonn, Germany

³Centre for Human Genetics, University of Marburg, Marburg, Germany

⁴Medical Faculty, Institute for Medical Biometry, Informatics and Epidemiology, University Bonn, Bonn, Germany

⁵Wellytics Technologies Pvt Ltd, Hyderabad, India

Received: 29 March 2023 / Accepted: 1 July 2023

Published online: 12 July 2023

References

- Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L et al. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res* 2022 Nov 9;51(D1):D977–85.
- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS Discovery: Biology, function, and translation. *Am J Hum Genet*. 2017 Jul;6(1):5–22.
- Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018 Sep;50(9):1219–24.
- Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, Neben CL et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun* 2020 Aug 20;11(1):3635.
- Hassanin E, May P, Aldisi R, Spier I, Forstner AJ, Nöthen MM et al. Breast and prostate cancer risk: The interplay of polygenic risk, rare pathogenic germline variants, and family history. *Genetics in Medicine*. 2022 Mar 1;24(3):576–85.
- Hassanin E, Spier I, Bobbili DR, Aldisi R, Klinkhammer H, David F, et al. Clinically relevant combined effect of polygenic background, rare pathogenic germline variants, and family history on colorectal cancer incidence. *BMC Med Genom*. 2023 Mar;5(1):42.
- Aldisi R, Hassanin E, Sivalingam S, Buness A, Klinkhammer H, Mayr A et al. GenRisk: a tool for comprehensive genetic risk modeling. *Bioinformatics*. 2022 May 1;38(9):2651–3.
- Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am J Hum Genet* 2022 Jan 6;109(1):12–23.
- Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet*. 2009;10:387–406.
- Wang M, Menon R, Mishra S, Patel AP, Chaffin M, Tanneeru D et al. Developing Genome-wide Polygenic Risk Scores for Coronary Artery Disease in South Asians. *J Am Coll Cardiol*. 2020 Aug 11;76(6):703–14.
- Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. *Nat Med*. 2022 Feb;28(2):243–50.
- Fritsche LG, Ma Y, Zhang D, Salvatore M, Lee S, Zhou X et al. On cross-ancestry cancer polygenic risk scores. *PLoS Genet* 2021 Sep 16;17(9):e1009670.
- Hodgson S, Huang QQ, Sallah N, Genes & Health Research Team, Griffiths CJ, Newman WG, et al. Integrating polygenic risk scores in the prediction of type 2 diabetes risk and subtypes in british Pakistanis and Bangladeshis: a population-based cohort study. *PLoS Med*. 2022 May;19(5):e1003981.
- Ho WK, Tan MM, Mavaddat N, Tai MC, Mariapun S, Li J et al. European polygenic risk score for prediction of breast cancer shows similar performance in asian women. *Nat Commun* 2020 Jul 31;11(1):3833.
- Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun* 2019 Jul 25;10:3328.
- Generalizing polygenic risk scores from Europeans to Hispanics/Latinos - Grinde - 2019 - Genetic Epidemiology - Wiley Online Library [Internet]. [cited 2022 Sep 29]. Available from: <https://onlinelibrary.wiley.com/doi/https://doi.org/10.1002/gepi.22166>
- Yang S, Zhou X. Accurate and scalable construction of polygenic scores in large Biobank Data Sets. *Am J Hum Genet* 2020 May 7;106(5):679–93.
- Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr Genet*. 2018 Oct;28(5):85–9.
- Wang M, Menon R, Mishra S, Patel AP, Chaffin M, Tanneeru D et al. Validation of a Genome-Wide Polygenic Score for Coronary Artery Disease in South Asians. *Journal of the American College of Cardiology*. 2020 Aug 11;76(6):703–14.
- Hao L, Kraft P, Berriz GF, Hynes ED, Koch C, Korategere V, Kumar P, et al. Development of a clinical polygenic risk score assay and reporting workflow. *Nat Med*. 2022 May;28(5):1006–13.
- Barnett AH, Dixon AN, Bellary S, Hanif MW, O'Hare JP, Raymond NT et al. Type 2 diabetes and cardiovascular risk in the UK south Asian community. *Diabetologia*. 2006 Oct 1;49(10):2234–46.
- Mars N, Lindbohm JV, della Briotta Parolo P, Widén E, Kaprio J, Palotie A et al. Systematic comparison of family history and polygenic risk across 24 common diseases. *The American Journal of Human Genetics*. 2022 Dec 1;109(12):2152–62.
- Hassanin E, Spier I, Bobbili DR, Aldisi R, Klinkhammer H, David F et al. Clinically relevant combined effect of polygenic background, rare pathogenic germline variants, and family history on colorectal cancer incidence [Internet]. medRxiv; 2022 [cited 2022 Sep 29]. p. 2022.01.20.22269585. Available from: <https://www.medrxiv.org/content/https://doi.org/10.1101/2022.01.20.22269585v1>
- Hassanin E, May P, Aldisi R, Krawitz P, Maj C, Bobbili DR. Assessing the role of polygenic background on the penetrance of monogenic forms in Parkinson's disease [Internet]. medRxiv; 2021 [cited 2022 Oct 6]. p. 2021.06.06.21253270. Available from: <https://www.medrxiv.org/content/https://doi.org/10.1101/2021.06.06.21253270v1>
- Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat Genet*. 2021 Apr;53(4):420–5.
- Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018 Oct;562(7726):203–9.
- Khera AV, Chaffin M, Wade KH, Zahid S, Brancale J, Xia R et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* 2019 Apr 18;177(3):587–596e9.
- Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*. 2019 Jul 1;8(7):giz082.
- Conroy MC, Lacey B, Bešević J, Omiyale W, Feng Q, Effingham M, et al. UK Biobank: a globally important resource for cancer research. *Br J Cancer*. 2023 Feb;128(4):519–27.
- Ekoru K, Adeyemo AA, Chen G, Doumatey AP, Zhou J, Bentley AR, et al. Genetic risk scores for cardiometabolic traits in sub-saharan african populations. *Int J Epidemiol*. 2021 Mar;17(4):1283–96.
- Graham SE, Clarke SL, Wu KHH, Kanoni S, Zajac GJM, Ramdas S, et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*. 2021 Dec;600(7890):675–9.
- Huang QQ, Sallah N, Dunca D, Trivedi B, Hunt KA, Hodgson S et al. Transferability of genetic loci and polygenic scores for cardiometabolic traits in british pakistani and bangladeshi individuals. *Nat Commun* 2022 Aug 9;13(1):4664.
- Hujoel MLA, Loh PR, Neale BM, Price AL. Incorporating family history of disease improves polygenic risk scores in diverse populations. *Cell Genom*. 2022 Jul;13(7):100152.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4. Discussion and Conclusion

In this thesis, an in-depth analysis of the UK Biobank cohort is conducted to investigate the interplay of genetic susceptibility, and family history in the prevalence and risk of breast, prostate, and colorectal cancers. The study uncovers the contributions of rare pathogenic variants, polygenic background, and family history to cancer risk, shedding light on the complex interplay between these factors and their implications for risk stratification and preventive measures.

However, one of the significant challenges in implementing PRS in clinical settings is their potential lack of transferability between different ancestral populations. Hence, we aim to assess the generalizability of PRS across different ethnic populations, including the South and East Asian populations, and to evaluate the potential challenges in implementing PRS in clinical settings due to population-specific differences.

4.1 Impact of polygenic risk scores

Recent studies have consistently demonstrated that the polygenic background, as represented by PRS, significantly modifies the risks for various cancers within the general population (Hsu et al. 2015; Mavaddat et al. 2019). Taking breast cancer as an example, a lower polygenic burden correspond with a substantial lower cancer risk (approximately one quarter of women participants on average), while a high PRS (above 80%) associated with a two-fold risk increase, and a very high PRS (99%) associated with almost quadruple cancer risk, reaching a magnitude comparable to that of carriers of hereditary breast cancer (Khera et al. 2018). Extending these findings, the present thesis utilizes the UK Biobank cohort, and the results confirm the strong modulation of breast cancer risk in the general population by the polygenic background. Specifically, individuals with low or high PRS exhibit a 0.5-fold or 2-fold change, respectively, in breast cancer odds compared to the average polygenic burden. Time-to-event analysis further reveals corresponding cumulative lifetime risks, highlighting the impact of PRS in identifying individuals at high risk. Additionally, the results underscore the heterogeneous nature of breast risk within patients with hereditary causes, emphasizing the need for a comprehensive understanding of genetic and polygenic influences in risk assessment (Lee et al. 2019).

4.1.1 Combining rare pathogenic variants, polygenic background, and family history

The findings reveal that both rare pathogenic variants and polygenic background factors contribute to the development of cancer. Specifically, the research suggests that individuals with suspected hereditary breast cancer are at a higher risk of developing breast cancer if they are carriers for rare pathogenic variants and have a higher PRS, those findings are consistent with other studies (Kuchenbaecker et al. 2017; Mavaddat et al. 2019; Mars et al. 2020b). However, there was not a significant interaction between rare pathogenic variant carrier status and PRS observed. The study also found that the lifetime risk of cancer is higher among individuals with rare pathogenic variants and a high PRS, highlighting the joint impact of both factors on cumulative disease incidence over an individual's lifetime.

The single-gene analysis revealed heterogeneous effects across genes, suggesting that the impact of PRS should be considered when assessing the absolute risk associated with individual genes. The findings suggest that incorporating PRS into risk stratification may help prevent unnecessary surveillance for breast cancer in individuals with moderate-risk genes, such as *CHEK2*, *PALB2* and *ATM*, while high-risk genes like *BRCA1/2* remain clinically relevant irrespective of PRS (Gallagher et al. 2020, 2021). Specially women with rare pathogenic variants in those moderate-risk genes and high PRS have a cumulative incidence of cancer that is comparable to that of women with rare pathogenic variants in high-risk genes (*BRCA1/2*) and low PRS. These findings support the inclusion of PRS in healthcare prevention policies, as it can identify a significant portion of the general population with risks comparable to those of rare pathogenic variants, particularly for moderate-risk genes (Kuchenbaecker et al. 2017). Additionally, further incorporating family history into risk stratification models improves predictive accuracy, particularly when combined with genetic factors (Plym et al. 2022; Ho et al. 2023). This comprehensive approach outperformed the consideration of a single risk factor, highlighting the importance of accounting for both genetic and familial elements in predictive models (Gao et al. 2021). Moreover, the added impact of family history may arise from unidentified genetic variants, including copy number variations, as well as non-genetic factors such as environmental and lifestyle influences.

4.2 Generalizability of European polygenic risk scores to South Asians

In the third study of the thesis, we examined the utility of European-based PRSs in the South Asian population, building on previous studies (Wang et al. 2020; Graham et al. 2021; Huang et al. 2022). The focus was on assessing how well PRSs from individuals with European heritage could predict CAD, T2D, and obesity in South Asians. Using UK Biobank data, our results suggest that a PRS from a large study of cardiometabolic diseases in Europeans could potentially help identify higher disease risks in South Asians. Nonetheless, it's important to note the decreased accuracy compared to the European group.

Our study involved adjusting for population structure to account for differences in genetic variants among ethnic populations (Privé et al. 2022), aiming to improve the accuracy of PRS predictions and their validation in diverse populations (Wang et al. 2020; Hao et al. 2022). While EUR-derived PRSs may have value in assessing risks in non-European populations, caution is needed due to recruitment limitations in the UK Biobank cohort. Variations in healthcare access, lifestyle, and non-genetic factors among ethnic groups highlight the importance of careful interpretation of results (Graham et al. 2021).

This study also looked at the combined effect of PRS and family history to evaluate disease risk in South Asians. The results indicate that combining PRS and family history could provide a more comprehensive understanding of disease risk, that considering both genetic and environmental factors (Mars et al. 2022). This is important for traits such as CAD, T2D, and obesity, especially in South Asians, where family environments and lifestyles can play a major role in disease risk. However, the success of this approach majorly depends on the disease and population being studied.

4.3 Clinical utility and challenges of PRS

PRSs have been extensively studied and demonstrated strong associations with disease risk. PRS may be used in modelling genetic risk and estimating life span an individual's disease risk (Lewis and Vassos 2020). PRS may play a role in therapeutic interventions either to prevent developing disease, like the use of statins in atherosclerosis prevention (Natarajan et al. 2017), or treating disease, like the use of

specific pathway PRS to augment chemotherapy dose for certain Leukemia patients (Elsayed et al. 2022). Many companies have been recently providing packages of informed life planning based on PRS, in the context of adopting lifestyle habits, recommendations for nutrition and sport (Torkamani et al. 2018).

However, the clinical utility of PRS is challenging and not yet established. There are number of challenges and shortcomings of translating PRS to clinics, ranging from ethical, technical, and practical concerns (Wang et al. 2022). The poor generalizability of PRS and attenuation of performance across diverse ancestries and cohorts is a significant challenge, which limit their clinical applicability (Peterson et al. 2019). This is obviously due to majority of genetic data sets are European-based cohorts, which may increase healthcare disparities. This issue is not unique to PRS, clinical models for risk factors in other areas of medicine as well as genetics context can lead to over and underestimation the risk for certain populations.

One challenge also is the lack of consensus for clinical guidelines for PRS use of different diseases (Hao et al. 2022). PRSs do not provide absolute risk for developing a disease but rather a probability of disease risk and trait likelihood. This difference in risk interpretation can lead to confusion for healthcare providers and communication of such scores with patients. PRS might be useful as an adjunct tool for modelling disease risk, rather than a definitive diagnosis.

Moreover, using millions of SNPs in calculating PRS, even many of those SNPS have minimal impact, raises a number of questions about the meaningfulness of including those SNPs in modelling risk prediction (St. -Pierre et al. 2022). Many studies showed that the performance achieved using a smaller set of significant genome-wide variants did not change significantly by using thousands or millions of SNPs, as the contribution of each SNP to the overall risk is often small (Mavaddat et al. 2019). Lastly, the interpretation and communicating of PRS results can be challenging due to the number of SNPs involved, potentially leading to confusion regarding the actual genetic predisposition to diseases. The use of a smaller set of genome-wide significant variants may provide a more manageable and interpretable risk prediction tool for both clinicians and patients.

4.4 Limitations

Limitations in our study include the selection bias towards "healthy volunteers" in the UK Biobank cohort, suggesting that the findings may not be fully applicable in terms of effect sizes (Fry et al. 2017). Furthermore, our risk assessment was based solely on genetic variations and family history, disregarding other established risk factors (Al Ajmi et al. 2020). For example, lifestyle-related elements, which have been highlighted as important in earlier UK Biobank studies, could have a substantial impact on cancer rates, while familial shared behaviors may affect the accuracy of reported family medical history (Kachuri et al. 2020). The impact of how common variants possibly interacting with rare variants on disease risk etiology remains an unanswered question. Lastly, while we examined the entire UK biobank cohort, the ability to apply risk stratification across diverse populations was hindered by the small sample size. The polygenic risk scores (PRS) may be skewed towards individuals of European descent, potentially reducing their predictiveness in non-European or mixed ancestry individuals, as shown in previous research (Fatumo et al. 2022; Privé et al. 2022).

4.5 Future work

We plan to extend this analysis of combining common and rare variants for other diseases, specially Parkinson's disease and epilepsy. The aim is to investigate the combined contribution of both common and ultra-rare genetic variants (URV) on the development of genetic generalized epilepsy (GGE) and GGE-sub phenotypes. A cohort with genotyping and whole-exome sequencing data from individuals with epilepsy and ancestry-matched controls will be analyzed. We plan to compute an individual risk score (IRS) that combines both common and rare variants risk scores. For rare risk score, the aim is to employ different allele frequency cutoffs and functional genomic annotations, to evaluate the impact of the IRS on epilepsy stratification within genes associated with GGE and GGE-sub phenotypes. The study will investigate the potential of using common and rare genetic variations to improve the classification and stratification of individuals with epilepsy.

Another study will involve the use of the latest epilepsy ILAE GWAS data (Stevelling et al. 2023), along with the available genotype cohorts from ILAE genetics and Epi25, and

defined gene-sets (Koko et al. 2021). This project will employ two approaches for polygenic assessment: pathway-PRS and multi-PRS. Pathway-specific PRSs will be utilized to provide more biological insights based on the genetic variants involved in specific epilepsy-related biological pathways (Choi et al. 2023). On the other hand, the multi-PRS framework will be used to generate many PRSs from publicly available GWAS data. As increasing the sample size for a specific phenotype is expensive and takes time. Multi-PRS approach can effectively increase the sample size by using genetically correlated phenotypes. The multi-PRS has demonstrated increased prediction accuracy over single PRS by exploiting the joint power of multiple discovery GWASs (Krapohl et al. 2018; Albiñana et al. 2023).

4.6 Conclusion

In conclusion, this thesis makes a significant contribution to our understanding of the complex factors impacting cancer susceptibility. It supports the careful inclusion of PRS in risk assessment models, taking into account familial, other genetic and non-genetic components. These results support the inclusion of PRS in Breast cancer prediction tools such as BODICEA and pave the way for its future integration in other cancers like prostate and colorectal cancer. The findings presented in the thesis underscore and showcase the clinical relevance of enhancing risk prediction approaches for personalized prevention and management strategies across different cancer types. Besides, personalizing PRSs based on specific ancestral backgrounds is essential for enhancing precision medicine.

5. References

Al Ajmi K, Lophatananon A, Mekli K, Ollier W, Muir KR. Association of Nongenetic Factors With Breast Cancer Risk in Genetically Predisposed Groups of Women in the UK Biobank Cohort. *JAMA Network Open*. 2020 Apr 24;3(4):e203760.

Albiñana C, Zhu Z, Schork AJ, Ingason A, Aschard H, Brikell I, et al. Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nat Commun*. 2023 Aug 5;14:4702.

Bhérer C, Eveleigh R, Trajanoska K, St-Cyr J, Paccard A, Nadukkalam Ravindran P, et al. A cost-effective sequencing method for genetic studies combining high-depth whole exome and low-depth whole genome. *npj Genom Med*. 2024 Feb 7;9(1):1–12.

Breast Cancer Association Consortium. Pathology of Tumors Associated With Pathogenic Germline Variants in 9 Breast Cancer Susceptibility Genes. *JAMA Oncology*. 2022 Mar 17;8(3):e216744.

Choi SW, García-González J, Ruan Y, Wu HM, Porras C, Johnson J, et al. PRSet: Pathway-based polygenic risk score analyses and software. *PLOS Genetics*. 2023 Feb 7;19(2):e1010624.

Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*. 2019 Jul 1;8(7):giz082.

Darst BF, Sheng X, Eeles RA, Kote-Jarai Z, Conti DV, Haiman CA. Combined Effect of a Polygenic Risk Score and Rare Genetic Variants on Prostate Cancer Risk. *European Urology*. 2021 Aug 1;80(2):134–8.

Dornbos P, Koesterer R, Rutenburg A, Nguyen T, Cole JB, Leong A, et al. A combined polygenic score of 21,293 rare and 22 common variants improves diabetes diagnosis based on hemoglobin A1C levels. *Nat Genet*. 2022 Nov;54(11):1609–14.

Duncan L, Shen H, Gelaye B, Meijssen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun*. 2019 Jul 25;10:3328.

Easton DF, Pharoah PDP, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, et al. Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *New England Journal of Medicine*. 2015 Jun 4;372(23):2243–57.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*. 2010 Jun;11(6):446–50.

Elsayed AH, Cao X, Mitra AK, Wu H, Raimondi S, Cogle C, et al. Polygenic Ara-C Response Score Identifies Pediatric Patients With Acute Myeloid Leukemia in Need of Chemotherapy Augmentation. *JCO*. 2022 Mar;40(7):772–83.

Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. *Nat Med*. 2022 Feb;28(2):243–50.

Fiziev PP, McRae J, Ulirsch JC, Dron JS, Hamp T, Yang Y, et al. Rare penetrant mutations confer severe risk of common diseases. *Science*. 2023 Jun 2;380(6648):eabo1131.

Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol*. 2017 Nov 1;186(9):1026–34.

Gallagher S, Hughes E, Kurian AW, Domchek SM, Garber J, Probst B, et al. Comprehensive Breast Cancer Risk Assessment for CHEK2 and ATM Pathogenic Variant Carriers Incorporating a Polygenic Risk Score and the Tyrer-Cuzick Model. *JCO Precis Oncol*. 2021 Jun 24;5:PO.20.00484.

Gallagher S, Hughes E, Wagner S, Tshiaba P, Rosenthal E, Roa BB, et al. Association of a Polygenic Risk Score With Breast Cancer Among Women Carriers of High- and Moderate-Risk Breast Cancer Genes. *JAMA Network Open*. 2020 Jul 1;3(7):e208501.

Gao C, Polley EC, Hart SN, Huang H, Hu C, Gnanaolivu R, et al. Risk of Breast Cancer Among Carriers of Pathogenic Variants in Breast Cancer Predisposition Genes Varies by Polygenic Risk Score. *JCO*. 2021 Aug 10;39(23):2564–73.

Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*. 2019 Apr 16;10(1):1776.

Ghose J, Sveinbjörnsson G, Vujkovic M, Seidelin AS, Gellert-Kristensen H, Ahlberg G, et al. Integrative common and rare variant analyses provide insights into the genetic architecture of liver cirrhosis. *Nat Genet*. 2024 Apr 17;1–11.

Gibson G. Rare and Common Variants: Twenty arguments. *Nat Rev Genet*. 2012 Jan 18;13(2):135–45.

Graham SE, Clarke SL, Wu KHH, Kanoni S, Zajac GJM, Ramdas S, et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*. 2021 Dec;600(7890):675–9.

Hao L, Kraft P, Berriz GF, Hynes ED, Koch C, Korategere V, Kumar P, et al. Development of a clinical polygenic risk score assay and reporting workflow. *Nat Med*. 2022 May;28(5):1006–13.

Ho PJ, Lim EH, Hartman M, Wong FY, Li J. Breast cancer risk stratification using genetic and non-genetic risk assessment tools for 246,142 women in the UK Biobank. *Genetics in Medicine*. 2023 Oct 1;25(10):100917.

Hsu L, Jeon J, Brenner H, Gruber SB, Schoen RE, Berndt SI, et al. A Model to Determine Colorectal Cancer Risk Using Common Genetic Susceptibility Loci. *Gastroenterology*. 2015 Jun 1;148(7):1330-1339.e14.

Huang QQ, Sallah N, Dunca D, Trivedi B, Hunt KA, Hodgson S, et al. Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals. *Nat Commun*. 2022 Aug 9;13(1):4664.

Hujoel MLA, Loh PR, Neale BM, Price AL. Incorporating family history of disease improves polygenic risk scores in diverse populations. *Cell Genom*. 2022 Jul 13;2(7):100152.

Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, et al. Principles and methods for transferring polygenic risk scores across global populations. *Nat Rev Genet*. 2024 Jan;25(1):8–25.

Kachuri L, Graff RE, Smith-Byrne K, Meyers TJ, Rashkin SR, Ziv E, et al. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat Commun*. 2020 Nov 27;11:6084.

Kessler MD, Damask A, O’Keeffe S, Banerjee N, Li D, Watanabe K, et al. Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature*. 2022;612(7939):301–9.

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018 Sep;50(9):1219–24.

Koko M, Krause R, Sander T, Bobbili DR, Nothnagel M, May P, et al. Distinct gene-set burden patterns underlie common generalized and focal epilepsies. *eBioMedicine*. 2021 Oct 1;72:103588.

Krapohl E, Patel H, Newhouse S, Curtis CJ, von Stumm S, Dale PS, et al. Multi-polygenic score approach to trait prediction. *Mol Psychiatry*. 2018 May;23(5):1368–74.

Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, et al. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA*. 2017 Jun 20;317(23):2402–16.

Kurki MI, Saarentaus E, Pietiläinen O, Gormley P, Lal D, Kerminen S, et al. Contribution of rare and common variants to intellectual disability in a sub-isolate of Northern Finland. *Nat Commun*. 2019 Jan 24;10:410.

Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in Medicine*. 2019 Aug 1;21(8):1708–18.

Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet*. 2014 Jul 3;95(1):5–23.

Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*. 2020 May 18;12(1):44.

Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, Lindbohm JV, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med*. 2020a Apr;26(4):549–57.

Mars N, Lindbohm JV, della Briotta Parolo P, Widén E, Kaprio J, Palotie A, et al. Systematic comparison of family history and polygenic risk across 24 common diseases. *The American Journal of Human Genetics*. 2022 Dec 1;109(12):2152–62.

Mars N, Widén E, Kerminen S, Meretoja T, Pirinen M, della Briotta Parolo P, et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat Commun*. 2020b Dec 14;11(1):6383.

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics*. 2017 Apr 6;100(4):635–49.

Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics*. 2019 Jan 3;104(1):21–34.

Natarajan P, Young R, Stitzel NO, Padmanabhan S, Baber U, Mehran R, et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*. 2017 May 30;135(22):2091–101.

Niemi MEK, Martin HC, Rice DL, Gallone G, Gordon S, Kelemen M, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*. 2018 Oct 1;562(7726):268–71.

Palshof FK, Mørch LS, Køster B, Engholm G, Storm HH, Andersson TML, et al. Non-preventable cases of breast, prostate, lung, and colorectal cancer in 2050 in an elimination scenario of modifiable risk factors. *Sci Rep*. 2024 Apr 13;14:8577.

Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*. 2019 Oct 17;179(3):589–603.

Plym A, Zhang Y, Stopsack KH, Jee YH, Wiklund F, Kibel AS, et al. Family History of Prostate and Breast Cancer Integrated with a Polygenic Risk Score Identifies Men at Highest Risk of Dying from Prostate Cancer before Age 75 Years. *Clin Cancer Res*. 2022 Nov 14;28(22):4926–33.

Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat Rev Genet*. 2019 Dec;20(12):747–59.

Price AL, Spencer CCA, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci*. 2015;282(1821):20151684.

Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*. 2022 Jan 6;109(1):12–23.

Rahim NG, Harismendy O, Topol EJ, Frazer KA. Genetic determinants of phenotypic diversity in humans. *Genome Biol*. 2008 Apr 24;9(4):215.

Roberts E, Howell S, Evans DG. Polygenic risk scores and breast cancer risk prediction. *Breast*. 2023 Jan 10;67:71–7.

Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*. 2023;73(1):17–48.

St. -Pierre J, Zhang X, Lu T, Jiang L, Loffree X, Wang L, et al. Considering strategies for SNP selection in genetic and polygenic risk scores. *Front Genet [Internet]*. 2022 Oct 25 [cited 2024 Apr 25];13. Available from: <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.900595/full>

Stevelling R, Campbell C, Chen S, Abou-Khalil B, Adesoji OM, Afawi Z, et al. GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture. *Nat Genet.* 2023 Sep;55(9):1471–82.

Susswein LR, Marshall ML, Nusbaum R, Vogel Postula KJ, Weissman SM, Yackowski L, et al. Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. *Genet Med.* 2016 Aug;18(8):823–32.

Tamlander M, Jermy B, Seppälä TT, Färkkilä M, Widén E, Ripatti S, et al. Genome-wide polygenic risk scores for colorectal cancer have implications for risk-based screening. *Br J Cancer.* 2024 Mar;130(4):651–9.

Toma C, Shaw AD, Allcock RJN, Heath A, Pierce KD, Mitchell PB, et al. An examination of multiple classes of rare variants in extended families with bipolar disorder. *Transl Psychiatry.* 2018 Mar 13;8:65.

Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet.* 2018 Sep;19(9):581–90.

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. *Nat Rev Methods Primers.* 2021 Aug 26;1(1):1–21.

Wang M, Menon R, Mishra S, Patel AP, Chaffin M, Tanneeru D, et al. Developing Genome-wide Polygenic Risk Scores for Coronary Artery Disease in South Asians. *J Am Coll Cardiol.* 2020 Aug 11;76(6):703–14.

Wang Y, Tsuo K, Kanai M, Neale BM, Martin AR. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annu Rev Biomed Data Sci.* 2022 Aug 10;5:293–320.

Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat Genet.* 2017 Jul;49(7):978–85.

List of Publications

1. Aldisi R, **Hassanin E**, Sivalingam S, Buness A, Klinkhammer H, Mayr A, et al. GenRisk: a tool for comprehensive genetic risk modeling. **Bioinformatics**. 2022 May 1;38(9):2651–3.
2. Aldisi R, **Hassanin E**, Sivalingam S, Buness A, Klinkhammer H, Mayr A, et al. Gene-based burden scores identify rare variant associations for 28 blood biomarkers. **BMC Genomic Data**. 2023 Sep 4;24(1):50.
3. Dueñas N, Klinkhammer H, Bonifaci N, Spier I, Mayr A, **Hassanin E**, et al. Ability of a polygenic risk score to refine colorectal cancer risk in Lynch syndrome. **Journal of Medical Genetics**. 2023 June 15;60:1044-1051.
4. **Hassanin E**, Lee KH, Hsieh TC, Aldisi R, Lee YL, Bobbili D, et al. Trans-ancestry polygenic models for the prediction of LDL blood levels: an analysis of the United Kingdom Biobank and Taiwan Biobank. **Frontiers in Genetics**. 2023 Nov 23;14:1286561.
5. **Hassanin E**, Maj C, Klinkhammer H, Krawitz P, May P, Bobbili DR. Assessing the performance of European-derived cardiometabolic polygenic risk scores in South-Asians and their interplay with family history. **BMC Medical Genomics**. 2023 Jul 12;16(1):164.
6. **Hassanin E**, May P, Aldisi R, Krawitz P, Maj C, Bobbili DR. Assessing the role of polygenic background on the penetrance of monogenic forms in Parkinson's disease. **medRxiv**. 2021. p. 2021.06.06.21253270.
7. **Hassanin E**, May P, Aldisi R, Spier I, Forstner AJ, Nöthen MM, et al. Breast and prostate cancer risk: The interplay of polygenic risk, rare pathogenic germline variants, and family history. **Genetics in Medicine**. 2022 Mar 1;24(3):576–85.
8. **Hassanin E**, Spier I, Bobbili DR, Aldisi R, Klinkhammer H, David F, et al. Clinically relevant combined effect of polygenic background, rare pathogenic germline variants, and family history on colorectal cancer incidence. **BMC Medical Genomics**. 2023 Mar 5;16(1):42.
9. Hess T, Maj C, Gehlen J, Borisov O, Haas SL, **Hassanin E**, et al. Dissecting the genetic heterogeneity of gastric cancer. **eBioMedicine**. 2023 Jun 1;92.
10. Stevelink R, Campbell C, Chen S, Abou-Khalil B, Adesoji OM, **Hassanin E**, et al. GWAS genetic architecture. **Nature Genetics**. 2023 Sep;55(9):1471–82.

6. Acknowledgements

The work in this thesis would not have been possible without the guidance and mentorship of all supervisors. I would like to express my sincere gratitude to my supervisor Prof. Peter Krawitz for the guidance, support, and mentorship provided. You have been very supportive throughout this PhD journey. Thank you for being flexible right from starting my PhD to the end. I would like to thank Dr. Patrick May for his time and patience during my PhD, especially time that I stayed in Luxembourg. You were always available whenever I needed you. You believed in my enthusiasm towards research and motivated me to push myself harder. Thanks a lot for your support after the transition from Bonn to Luxembourg.

Dr. Carlo Maj and Dr. Dheeraj Bobbili, I was extremely lucky having you by my side as experienced advisors. You make things much easier for me and provide support whenever I needed. Thank you for motivating me and be always on my side, supporting me throughout all my PhD. It was great to work with you on the PRS, GWAS and rare variants together and to be able to share and discuss every result within minutes.

I would like to thank Prof. Andreas Meyer, Prof. Markus Noten for accepting to be part of my thesis committee and for their insightful comments and suggestions that significantly enhanced the quality of this work. I would like to thank Prof. Dr. Stefan Aretz and Dr. Isabel Spier for the collaboration on colorectal cancer analysis. I would like also to thank Ko-Han Lee and Prof. Chien-Yu Chen for the collaboration on the study of PRS in Taiwan Biobank. Your contributions have enriched the intellectual environment of my research.

I would like to acknowledge all my fellow researchers and colleagues at IGSB University of Bonn and LCSB University of Luxembourg who have shared their knowledge, provided encouragement, and engaged in fruitful discussions. Especially, Rana Aldisi, Hannah Klinkhammer, Dr. Zied Landoulsi, and Maryam Erfanian Omidvar.

To all my good friends in Germany and Luxembourg, Abdullah, Hurya, Waheed, Soudy, Yaghmour, Hossam, Suzan, and Edi, I cannot thank you enough for always being there for me, you made my life in Germany and Luxembourg a lifelong great memory and less stressful. There are also many people in Luxembourg and Germany who helped me at various times, I owe them a lot. I thank all my friends back in Egypt, especially Salah,

Abdelrahman, and Refaat. Each one of you have made a difference in my life and I have no words to express my feelings towards you. You guys never forget me, and I am very lucky to have you as my friends.

I would like to deeply thank my family for their unwavering support, understanding, and encouragement. I believe this thesis is dedicated to my parents, Saeed and Mervat, and my brothers Abdo, Mohamed, my sister Eman, and my wife Aya. Your belief in me has been a constant source of motivation.