Genomic ascertainment of *CHEK2*-related cancer predisposition

Running head: *CHEK2*-related cancer predisposition

*Sun Young Kim, MD, Ph.D[1,2], *Jung Kim, Ph.D[1], Mark Ramos, Ph.D[1], Jeremy Haley, MS[3], Diane Smelser, Ph.D[3], H. Shanker Rao, MS[3], Uyenlinh L. Mirshahi, Ph.D[3], Geisinger-Regeneron DiscovEHR Collaboration, Barry I. Graubard, Ph.D[4], Hormuzd A. Katki, Ph.D[4], David Carey, Ph.D[3], Douglas R. Stewart, MD[1]

[1]Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Rockville, MD, USA, [2]Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA, [3] Department of Genomic Health, Geisinger, Danville, PA, USA, [4]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Rockville, MD, USA

*Sun Young Kim and *Jung Kim contributed to equally to this research.

Correspondence to: Douglas R. Stewart

Address: 9609 Medical Center Drive Rm 6E450, Bethesda, MD, 20892

Tel: 240-276-7238; Fax: 240-276-7836.

Email: drstewart@mail.nih.gov

Acknowledgments: This work was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics of the National Cancer Institute, Bethesda, MD and utilized the computational resources of the NIH High-Performance Computing Biowulf cluster. This research has been conducted using the UK Biobank Resources under application 54389.

Previous presentations: Portions of this investigation were presented as a platform presentation at the annual meeting of the American College of Medical Genetics and Genomics in Salt Lake City, Utah, March 2023.

Abstract

Purpose. There is clear evidence that deleterious germline variants in *CHEK2* increases risk for breast and prostate cancers; there is limited or conflicting evidence for other cancers. Genomic ascertainment was used to quantify cancer risk in *CHEK2* germline pathogenic variant heterozygotes.

Patients and Methods. Germline *CHEK2* variants were extracted from two exome-sequenced biobanks linked to the electronic health record: UK Biobank (n= 469,765**)** and Geisinger MyCode (n=170,503**)**. Variants were classified as per American College of Medical Genetics and Genomics (ACMG)/Association for Molecular Pathology (AMP) criteria. Heterozygotes harbored a *CHEK2* pathogenic/likely pathogenic (P/LP) variant; controls harbored benign/likely benign *CHEK2* variation or wildtype *CHEK2*. Tumor phenotype and demographic data were retrieved; to adjust for relatedness, association analysis was performed with SAIGE-GENE+ with Bonferroni correction.

Results. In *CHEK2* heterozygotes in both MyCode and UK Biobank, there was a significant excess risk of all cancers tested, including breast cancer (C50; OR=1.54 and 1.84, respectively), male genital organ cancer (C60-C63; OR=1.61 and 1.77 respectively), urinary tract cancer (C64-C68; OR=1.56 and 1.75, respectively) and lymphoid, hematopoietic, and related tissue cancer (C81-C96; OR=1.42 and 2.11, respectively). Compared to controls, age-dependent cancer penetrance in *CHEK2* heterozygotes was significantly younger in both cohorts; no significant difference was observed between the penetrance of truncating and missense variants for cancer in either cohort. Overall survival was significantly decreased in *CHEK2* heterozygotes in UK Biobank but there was no statistical difference in MyCode.

Conclusion. Using genomic ascertainment in two population-scale cohorts, this investigation quantified the prevalence, penetrance, cancer phenotype and survival in *CHEK2* heterozygotes. Tailored treatment options and surveillance strategies to manage those risks are warranted.

Introduction

CHEK2 (OMIM 604373) is a tumor-suppressor gene that encodes CHK2 (Serine/threonine protein kinase), involved in DNA repair in response to cellular DNA damage.[1] There is clear evidence that deleterious germline variants in *CHEK2* heterozygotes are associated with an increased risk for female breast and prostate cancers; however, elevated risks for a variety of other cancers (e.g., colorectal, kidney, bladder, leukemia/lymphoma and thyroid) have been claimed but there is minimal, biased or conflicting evidence.[2] In general, germline pathogenic truncating variants (PTV) (*e.g.*, c.1100del p.(Thr367fs)) are associated with an increased risk of cancer. In contrast to PTV, pathogenic missense variants (PMV) in *CHEK2* have more variable effects, mainly dependent on whether a critical protein domain is affected. According to a study by Dorling et al.[3], approximately 60% of rare PMV in *CHEK2* are associated with a lower risk of developing cancers compared to PTV. This suggests that the impact of PMV on cancer susceptibility is not uniform, but rather depends on the specific location and nature of the variants. To date, most work on quantifying risk from a germline variant in a cancer-predisposition gene has arisen from the well-established phenotype-first approach, in which individuals (and families) are ascertained from their presentation due to a clinical problem.

Genomic ascertainment is the inversion of the traditional phenotype-first approach[4]. With genomic ascertainment, germline variation of interest is identified, and phenotype status is then obtained from medical records to estimate variant prevalence and disease penetrance and characterize the phenotype. In principle, this should permit a less-biased estimate of the phenotypic spectrum, expressivity and penetrance of a deleterious variant or set of variants. Ascertainment biases still exist depending on how the cohort was recruited (healthier volunteer vs. clinical (health system or hospital) vs. true population sampling) that will influence risk estimates.

In this study, we used genomic ascertainment to quantify cancer risk for heterozygotes with germline pathogenic *CHEK2* variants. We analyzed electronic health record (EHR) in two population-based

cohorts (UK Biobank (UKBB) and Geisinger MyCode) to estimate the prevalence, age-dependent penetrance, cancer risk and survival of *CHEK2* pathogenic heterozygotes compared to controls.

Materials and Methods

Cohorts and relatedness

From the UK Biobank, germline variants were obtained from field 23157, population level exome OQFE variants, and pVCF format, final exome release. Human subjects' protection and review was through the North West Multi-centre Research Ethics Committee. Exome sequencing on UKBB samples has been described.[5,6] The data was accessed January 2023; the number of unrelated participants was determined by R package ukbtools, "ukb_gene_samples_to_remove" function.

Geisinger is an integrated health system serving patients in Northeastern and Central Pennsylvania. All Geisinger patients are eligible to participate in the MyCode Community Health Initiative, a system-wide biorepository of blood and DNA samples for broad research purposes.[7] Over 85% of Geisinger patients agree to participate and provided genomic data that are linked to their health records, which consist of routinely collected diagnosis, procedures, medication, and laboratory results, collected as part of their healthcare. This study was approved by the Geisinger Institutional Review Board. MyCode DNA samples were exome sequenced by the Regeneron Genetics Center using IDT exon capture probes as previously described.[8] This study was approved by the Geisinger Institutional Review Board. In the Geisinger MyCode cohort, we included individuals over 18 years of age (n=167,050 with available exome data). To remove related individuals while maintaining the largest possible cohort, kinship pairs up to 3rd degree relatives (minimum PI_HAT = 0.1875) were used to create a graph of all relatives. Custom functions employing the network library in python were used. For each connected component (i.e., family), the node (i.e., patient) with the greatest number of edges (i.e., relatives) was removed. This was repeated until no edges remain in the connected component.

Variant filtering and CHEK2 pathogenicity classification

Variants were filtered on the following quality metrics: Allelic Balance of Heterozygotes (ABHet) between 0.2 and 0.8, Genotype Quality >30, total read depth>5. All variants that pass quality metrics were annotated using snpEFF[9], ANNOVAR[10], ClinVar[11] (database retrieved 09-23-2022), and InterVar (v.2.1.3)[12]. Variants were classified as pathogenic (P), likely pathogenic (LP), variant of uncertain significance (VUS), likely benign (LB), benign (B) using guidelines from the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP).[13] Final variant annotation was based on a hierarchical classification of ClinVar followed by InterVar[14]. "Heterozygotes" were defined as individuals who harbor a *CHEK2* P/LP variant, whereas controls included individuals who harbor canonical or B/LB *CHEK2* variation. *CHEK2* VUS were excluded. In this analysis, "All" refers to all *CHEK2* P/LP variants, "PTV" refers to predicted *CHEK2* truncating P/LP variants and "PMV" refers to pathogenic missense *CHEK2* P/LP variants. There were eight individuals and six individuals who harbored biallelic *CHEK2* variants in UKBB and MyCode, respectively; they were included in the All group, but were excluded from analyses of PTV and PMV. There was no individual who carried more than two P/LP variants in either UKBB or MyCode.

Cancer phenotype and vital status query

Tumor phenotype and demographic data (age, sex, body mass index (BMI), alcohol consumption, smoking history, and race) were obtained for both heterozygotes and controls. Demographic comparisons were completed using Student T-test for continuous variables and Fisher's exact test for binary variable. Clinical phenotypes of neoplasms were obtained using International Classification Diseases (ICD) diagnosis codes: ICD9 and ICD10 for UKBB, ICD10-Clinical Modification (CM) for MyCode data. The Geisinger Cancer Registry was also queried, which contains information on all patients diagnosed with cancer at a Geisinger facility; the Cancer Registry (field 40006 and 40013) and Death Registry data for UKBB (field 40001) were also utilized.

Power to detect predisposition to common and rare cancers in UK Biobank and MyCode

Power estimates were performed by adapting formulas from Chow et al.[15] to a cohort study setting with the assumption of non-biased ascertainment.

Supplemental Figure 1 shows power as a function of presumed true odds ratio for a range of cancer rates in the UK Biobank and MyCode cohorts using cohort-specific All, PTV and PMV *CHEK2* heterozygote prevalence from **Table 1**. For All, PTV and PMV *CHEK2* heterozygotes, there is 100% power in both UK Biobank and MyCode to detect common cancers (≥5% cancer rate, which include many sex-specific cancers such as female breast and prostate) with an odds ratio of >2. For All, PTV and PMV *CHEK2* heterozygotes, there is ≥80% power to detect rare cancers (≥1% cancer rate) with an odds ratio of >2. For All *CHEK2* heterozygotes, there is >80% power to detect very rare cancers (≥0.1%) with an odds ratio of >2.7 in both MyCode and UKBB; there is less power in the PTV- and PMV-specific cohorts.

Cancer risk estimate

Cancer prevalence was modeled using logistic regression with carrier status for All, PTV, and PMV as the main set of explanatory variables and age, sex, smoking history, alcohol consumption and BMI as covariates. For sex-specific cancers (C51-C58 for female; C60-C63 for male) prevalence was analyzed only with female or male controls. Multiplicity issues were addressed using Bonferroni adjustment at family-wise error rate of α=0.05. To correct for relatedness, we used SAIGE-GENE+ version 1.1.6.2. Covariates include using PC1-4, current age, sex, smoking, alcohol use and BMI.[16] To further help guard against inaccurate p-values and confidence interval coverage for associations arising from very low prevalence of rare cancers, cancers where there were less than five cases among heterozygotes were excluded from any analyses.

Kaplan-Meier, cancer penetrance and mortality in individuals with cancer

Kaplan-Meier survival analyses were used to estimate all-cause mortality, penetrance of pathogenic *CHEK2* variants for cancer, and overall survival for individuals with cancer in MyCode and UKBB

cohorts. In the MyCode cohort, only events occurring 3 months or after in Geisinger facilities are included in survival analyses. Data was truncated to individuals with current age $\leq$ 85. Hazard ratios were computed using Cox Proportional-Hazards (coxph) model adjusting for age, self-reported race, sex, smoking history, alcohol consumption and BMI, using Log-rank test for equality to compare differences between the curves for controls and variant groups. Coxph also adjusted for relatedness by clustering genetically inferred family units. All the analyses were conducted using R version 4.1.2.

Results

Prevalence and demographics of All, PTV and PMV CHEK2 heterozygotes in MyCode and UK Biobank

Table 1 shows the prevalence of All, PTV and PMV *CHEK2* heterozygotes in both cohorts. (The sum of PTV and PMV is less than All total due to the presence of non-canonical splice-site variants; **Supplemental Table 1** provides details on the variants) The relatedness (up to the third degree) of the MyCode and UKBB cohorts is ~30% and ~10% respectively; **Table 1** also shows the heterozygote prevalence in the unrelated fraction of the two cohorts. **Supplemental Table 2** lists demographics and covariates between All, PTV and PMV heterozygotes and controls. There were 305,330 controls (65%) in UKBB and 152,662 controls (91%) in MyCode.

Significant excess risk for cancers of the breast, male genital organ, urinary tract and lymphoid, hematopoietic, and related tissues in CHEK2 heterozygotes in both MyCode and UKBB.

Figure 1A displays statistically significant association of pathogenic *CHEK2* All, PTV and PMV for organ system groupings of cancer in MyCode. The odds ratios and Bonferroni-corrected p-values for the association between *CHEK2* heterozygotes for organ system groupings of cancer ICD codes are shown. In All *CHEK2* heterozygotes, there was a significant excess risk (Bonferroni-adjusted SAIGE p-value) of all cancers, breast cancer (C50), male genital organ cancer (C60-C63), urinary tract cancer (C64-C68), thyroid and other endocrine gland cancer (C73-C75), and lymphoid, hematopoietic, and related tissue cancer (C81-C96). (Of all the C50 codes observed in *CHEK2* heterozygotes, 155 and 225 (99.4, 99.1%)

were in females and, 1 and 2 (0.6, 0.8%) were in males in MyCode and UK Biobank, respectively.) **Supplemental Figure 2** displays the odds ratio for the MyCode cohort for All, PTV and PMV *CHEK2* heterozygotes for all organ system groupings of cancer ICD codes. **Figure 1B** displays the odds ratio for the UKBB cohort for All, PTV and PMV *CHEK2* heterozygotes for organ system groupings of cancer ICD codes with a significant excess of risk. In All *CHEK2* heterozygotes, there was a significant excess risk (Bonferroni-adjusted SAIGE p-value) of developing all cancers, breast cancer (C50), male genital organ cancer (C60-C63), urinary tract cancer (C64-C68), cancer from the secondary and unspecified sites (C76-C79), and lymphoid, hematopoietic, and related tissue cancer (C81-C96). In contrast to MyCode, there was a non-significant excess risk to develop thyroid and other endocrine gland cancer (C73-C75). **Supplemental Figure 3** displays the odds ratio for the UKBB cohort for All, PTV and PMV *CHEK2* heterozygotes for all organ system groupings of cancer ICD codes.

Specific cancers related to All, PTV and PMV CHEK2 heterozygotes

Figure 2A shows the specific types of cancer in the MyCode cohort with an excess risk from the significant organ-system analysis shown in **Figure 1A**. Of note is the significant excess risk for prostate cancer (C61), kidney cancer (C64), bladder cancer (C67), thyroid cancer (C73), and lymphoid leukemia (C91) in All heterozygotes. **Supplemental Figure 4** displays the odds ratio for the MyCode cohort for All, PTV and PMV *CHEK2* heterozygotes for all specific types of cancer from all organ system groupings of cancer ICD codes. **Supplemental Table 3** lists the case counts and percentages for PMV, PTV and All cohorts and fold-enrichment (vs. controls) for each of the ICD10 diagnostic codes in MyCode.

Figure 2B shows the specific types of cancer in the UKBB cohort with an excess risk from the organ-system analysis shown in **Figure 1B**. Of note is the significant excess risk for prostate (C61), kidney cancer (C64), and bladder cancer (C67) in All and PTV heterozygotes. There was significant increased risk for diffuse non-Hodgkin lymphoma (C83), other and non-specified types of non-Hodgkin lymphoma

(C85) and lymphoid leukemia (C91) in All heterozygotes, whereas peripheral and cutaneous T-cell lymphomas (C84) were exclusively associated with PMV heterozygotes. **Supplemental Figure 5** displays the odds ratio for the UK Biobank cohort for All, PTV and PMV *CHEK2* heterozygotes for all specific types of cancer from all organ system groupings of cancer ICD codes. **Supplemental Table 3** lists the case counts and percentages for PMV, PTV and All cohorts and fold-enrichment (vs. controls) for each of the ICD10 diagnostic codes in UK Biobank.

Age-dependent penetrance differs significantly in All, PTV and PMV CHEK2 heterozygotes vs. controls, but not between CHEK2 PTV vs. PMV heterozygotes

Compared to controls, age-dependent penetrance in All *CHEK2* variants for all cancers was significantly different in both MyCode (adjusted HR: 1.26 [95%CI 1.17-1.36], P-value: 6.1×10^{-10}) and UKBB (adjusted HR 1.31 [95%CI 1.24-1.40], *P*-value: 2.0×10^{-16}) (**Figures 3A and 4A**). *CHEK2* PMV or PTV heterozygotes alone were at higher risk for all cancers tested compared to controls in both MyCode and UKBB. In MyCode, for PMV (vs. controls) the adjusted HR: 1.24 [1.13-1.35], p value=2.71×10^{-06}; in UKBB, for PMV (vs. controls) the adjusted HR: 1.17 [1.06-1.30], p-value=1.56×10^{-3}. In MyCode, for PTV (vs. controls) adjusted HR: 1.30 [1.13-1.50], p value=2.1×10^{-4}; in UKBB for PTV (vs. controls) the adjusted HR:1.34 [1.23-1.45], p-value=1.67×10^{-12}. There was no significant difference in the penetrance of *CHEK2* PTV vs. PMV for cancers in MyCode (univariate HR: 1.06 [0.90-1.25], *P*-value=0.47) and the UKBB (adjusted HR 1.15 [95%CI 1.00-1.33], *P*-value: 0.05).

All-cause mortality was significantly increased in All heterozygotes compared to controls in UKBB but not MyCode

All-cause mortality was significantly increased in All heterozygotes in UKBB (adjusted HR 1.21 [95%CI 1.08-1.37], P-value: 1.51E-3) but not in MyCode (adjusted HR 1.09 [95%CI 0.96-1.24], *P*-value: 0.20) (**Figures 3B and 4B**). There was no significant difference in all-cause mortality in PTV and PMV

heterozygotes in UKBB (adjusted HR 1.24 [95%CI 0.97-1.60], *P*-value: 0.10) and MyCode (adjusted HR 1.06 [95%CI 0.80-1.41], *P*-value: 0.67).

All-cause mortality amongst individuals with cancer was not significantly increased in All heterozygotes compared to controls in UKBB and MyCode

There was no statistical difference between All heterozygotes and controls in both MyCode (adjusted HR 1.08 [95%CI 0.90-1.30], *P*-value=0.43) and the UKBB (adjusted HR 1.12 [95%CI 0.98-1.29], *P*-value: 0.11) cohorts for all-cause mortality in individuals with cancer. There were no significant differences between PTV and PMV heterozygotes in either the MyCode (adjusted HR 1.20 [95%CI 0.81-1.78], *P*-value=0.35) or UKBB (adjusted HR 1.33 [95%CI 0.98-1.80], *P*-value: 0.07) cohorts (**Figure 3C and 4C**).

Discussion

In this investigation, familial relationship-adjusted, Bonferroni-corrected genomic ascertainment of two population-based, exome-sequenced, EHR-linked cohorts was used to quantify risk of cancers arising from pathogenic/likely pathogenic germline variants in *CHEK2*. Notably, given the stated assumptions about participant ascertainment, both cohorts had high power to detect elevated risk (OR>2) in all but the rarest cancers. Genomic ascertainment quantifies risk based on genotype (not phenotype) and thus may reduce risk inflation arising from cancer ascertainment (case/family recruitment) by personal and/or family medical history.

Clinically, this investigation confirms the significantly increased risk for breast and prostate cancers (as well as all cancers, collectively), although the observed risk tends to be even lower (OR<2) than previous estimates, especially for PTV (typically OR>2).[2] Interestingly, in neither cohort was a significant excess risk for "malignant neoplasms of digestive organs" (majority were colorectal cancers) observed for All, PTV or PMV (**Supplemental Table 3**). Published risk estimates for colorectal cancer from *CHEK2* PTV are more modest (OR ~2) and conflicting than those for female breast cancer and prostate cancer; higher estimates of risk are driven by studies of multiplex families.[17] Published risk for colorectal cancer from

CHEK2 PMV tend to be even lower (OR<2) or non-significant.[2,18] Given this, a recent ACMG review and clinical practice guideline on management[2] concluded that *CHEK2* heterozygosity is not clinically actionable for colorectal cancer risk in isolation and to offer surveillance as per family history. In contrast, current National Comprehensive Cancer Network (NCCN) guidelines recommend colorectal cancer screening for individuals who carry *CHEK2* P/LP variants. (https://www.nccn.org/professionals/physician_gls/pdf/genetics_colon.pdf) Our observations in this study of non-significant colorectal risk are congruent with the ACMG recommendations. In summary, although additional confirmation is needed for breast, prostate and colorectal cancers, genomic ascertainment showed generally lower (or non-significant) risk than previously reported for All, PTV and PMV in *CHEK2*.

This work provides substantial evidence from both cohorts of significant increased risk for kidney cancer, bladder cancer and CLL (lymphoid leukemia). In this investigation, Bonferroni correction was applied to organ-system groupings and not specific cancer types. Thus, other cancers may be enriched in *CHEK2* heterozygotes; **Supplemental Table 2** lists counts of cancer types in controls and All, PTV and PMV heterozygotes. Recent ACMG clinical practice guideline on management of *CHEK2* heterozygotes[2] concluded that there was likely an increased risk for kidney cancer but that larger studies with appropriate controls were needed. Several publications found a range of risk (OR=3; hazard ratio =10.8)[18-21]; other investigations had non-significant findings.[22] As with breast and prostate cancers in this study, genomic ascertainment resulted in lower risk estimates (OR<2) for kidney cancer than previous studies and was remarkably consistent across the two cohorts. A 2023 ACMG review and clinical guidance for *CHEK2* heterozygotes[2] noted a single publication of non-significant *CHEK2*-associated bladder cancer[23] but deemed this evidence insufficient to make recommendations; more recent publications have found additional evidence of a *CHEK2*-bladder cancer association.[24,25] Genomic ascertainment in this study revealed similarly increased bladder cancer risk in both cohorts (especially in PTV). In summary, this

evidence of increased risk for both renal and bladder cancers should prompt the development of clinical management recommendations for surveillance and intervention for these cancers.

In both cohorts there was significantly elevated risk for lymphoid and hematopoietic neoplasms collectively (C81-C96); across all the subtypes of these malignancies, only CLL (lymphoid leukemia) had significantly elevated risk (>2) in both cohorts. Reports of increased risk of hematologic malignancy (especially CLL) in *CHEK2* heterozygotes date from 2006[26,27] but were conflicting and/or based on highly ascertained families. A 2022 investigation using a PheWAS approach in an earlier version of UKBB reported an excess risk (OR>3) for leukemia and plasma cell neoplasms in *CHEK2* P/LP heterozygotes.[28] To date with current approaches (*e.g.,* CBC, physical exam) there is limited evidence-based actionability for surveillance for increased risk of leukemia, however with developing methods (*e.g.,* methylation profiling of circulating tumor DNA) this may improve. Outcomes and tailored treatment options for *CHEK2*-associated CLL merit investigation.

A significant excess of malignancies of thyroid and other endocrine tumors (C73-C75) was observed in MyCode but not UK Biobank; this was almost entirely driven by thyroid tumors (C73) and, unlike most other associations, by *CHEK2* PMV. Previous studies have been conflicting or limited by small numbers or single-country ascertainment.[18,22,29] The recent ACMG review and clinical guidance for *CHEK2* heterozygotes[2] did not find sufficient evidence to support a clear association for thyroid cancer and did not recommend surveillance. Genomic ascertainment of *DICER1*-associated thyroid disease (*e.g.,* goiter) also found significant differences in *DICER1* heterozygotes (vs. controls) in MyCode but not UK Biobank and may reflect the different medical cultures in the US and UK in approaches to medical imaging of the thyroid.[30] Conversely, there was a significant excess risk of "malignant neoplasms of ill-defined, secondary and unspecified sites" (C76-C79) in UK Biobank but not MyCode.

Numerous other associations have been observed for specific cancers for *CHEK2* heterozygotes including sarcoma, stomach, male breast, melanoma, endometrial and testicular cancer[2]. For more common cancers

(endometrial, skin), there was no evidence of association for these in either cohort. For some rarer cancers (male breast, testicular) the two cohorts were likely underpowered (**Supplemental Figure 1**); for others (sarcoma, stomach) there may be both a power issue and a survival bias in ascertainment given the aggressive nature of these cancers.

Overall, pathogenic germline *CHEK2* All, PTV and PMV are common, but the conferred excess cancer risk is, with few exceptions, less than an OR of 2. In addition, the lack of significant difference between *CHEK2* All heterozygotes and controls in all-cause mortality in individuals with cancer suggests that germline *CHEK2*-associated cancer is not clinically more aggressive than non-*CHEK2*-associated cancer. The degree of risk from PTV and PMV overlap considerably with risk of PMV generally lower. The clinical relevance of this may be debatable since penetrance for cancer, all-cause mortality and all-cause mortality in individuals with cancer was not significantly different between PMV and PTV in both cohorts.

There are limitations to these retrospective analyses. MyCode and UK Biobank are predominantly of European ancestry. Copy-number (deletions) in *CHEK2* were not evaluated due to limited data availability in UK Biobank. Enrollment in the two cohorts was subject to ascertainment biases as individuals with conditions leading to death or disabilities would be less likely to participate. The "healthy volunteer" bias (compared to the UK population) of the UK Biobank has been documented [31]

In summary, we quantified cancer risk and survival in *CHEK2* heterozygotes using the novel genome-first approach in two well-powered cohorts. Our findings inform clinical care by supporting current recommendations for prostate and breast cancer surveillance and provide definitive evidence of increased risk for renal, bladder, and CLL in heterozygotes with pathogenic *CHEK2* variants. Tailored treatment options and surveillance strategies to manage those risks are needed.

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government.

The authors would like to acknowledge the participants of the MyCode Community Initiative for the use of their health and genomic information, without whom this study would not be possible. The patient enrollment and exome sequencing were funded by the Regeneron Genetics Center. Data for this project was made possible by the Geisinger-Regeneron DiscovEHR Collaboration.

Figures

Figure 1. Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for organ system groupings of cancer ICD codes with a significant excess of risk in MyCode (panel A) and UK Biobank (panel B). CI: 95% confidence interval; OR: odds ratio; PMV: pathogenic missense variant; PTV: pathogenic truncating variant

Figure 2. Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for specific cancers in the organ system groupings of cancer ICD codes with a significant excess of risk in MyCode (panel A) and UK Biobank (panel B). CI: 95% confidence interval; OR: odds ratio; PMV: pathogenic missense variant; PTV: pathogenic truncating variant

Figure 3. Penetrance of pathogenic *CHEK2* variants for cancer and all-cause mortality in MyCode. **Panel A**: Time-to-cancer (penetrance); **Panel B**: All-cause mortality; **Panel C**: All-cause mortality for individuals with cancer. PMV: pathogenic missense variant; PTV: pathogenic truncating variant.

Figure 4. Penetrance of pathogenic *CHEK2* variants for cancer and all-cause mortality in UK Biobank. **Panel A**: Time-to-cancer (penetrance); **Panel B**: All-cause mortality; **Panel C**: All-cause mortality for individuals with cancer. PMV: pathogenic missense variant; PTV: pathogenic truncating variant

Supplemental Figures

Supplemental Figure 1. Power as a function of risk (odds ratio) in MyCode (**Panel A, C, E**) and UK Biobank (**Panels B, D, F**) for a range of cancer rates. Prevalence data from cohort-specific ALL (Panels A, B) pathogenic truncating variants (PTV) (Panels C, D) and pathogenic missense variants (PMV) (Panels E, F) *CHEK2* heterozygotes (Table 1). Dark gray line represents 80% power, and light gray line represents 90% power.

Supplemental Figure 2. Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for organ system groupings of cancer ICD codes in MyCode. Red font represents significant cancers. CI: 95% confidence interval; OR: odds ratio; PMV: pathogenic missense variant; PTV: pathogenic truncating variant

Supplemental Figure 3. Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for organ system groupings of cancer ICD codes in UK Biobank. Red font represents significant cancers. CI: 95% confidence interval; OR: odds ratio; PMV: pathogenic missense variant; PTV: pathogenic truncating variant

Supplemental Figure 4. Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for all specific cancers in the organ system groupings of cancer ICD codes in MyCode. Red font represents significant cancers. CI: 95% confidence interval; OR: odds ratio; PMV: pathogenic missense variant; PTV: pathogenic truncating variant

Supplemental Figure 5. Odds ratio for All, PTV and PMV *CHEK2* heterozygotes for all specific cancers in the organ system groupings of cancer ICD codes in UK Biobank. Red font represents significant cancers. CI: 95% confidence interval; OR: odds ratio; PMV: pathogenic missense variant; PTV: pathogenic truncating variant

Supplemental Tables

Supplemental Table 1. List of all variants found in the study

Supplemental Table 2. Demographics of *CHEK2* heterozygotes vs. controls

Supplemental Table 3. Case counts and percentages for PMV, PTV and All cohorts and fold-enrichment (vs. controls) for each of the ICD10 diagnostic codes in MyCode and UK Biobank.

Table 1. Prevalence of All, pathogenic truncating variants (PTV) and pathogenic missense variants (PMV) *CHEK2* in adult heterozygotes in UK Biobank and Geisinger MyCode. The sum of PTV and PMV is less than All total due to presence of non-canonical splice-site variants.

Cohort	Individuals/Prevalence (95%CI)	All *CHEK2* P/LP Variants	Pathogenic Truncating Variants (PTV)	Pathogenic Missense Variants (PMV)
UK Biobank – related and unrelated (n=469,765)	Number of individuals	3,232	1,847	1,290
	Prevalence	1/145 (1/140 – 1/150)	1/254 (1/243 – 1/266)	1/364 (1/344 – 1/384)
UK Biobank – unrelated (n=437,645)	Number of individuals	3,171	1,825	1,268
	Prevalence	1/138 (1/133-1/142)	1/239 (1/229-1/251)	1/345 (1/326-1/364)
MyCode – related and unrelated (n=167,050)	Number of individuals	3,153	913	2,221
	Prevalence	1/52 (1/51 – 1/54)	1/183 (1/171 – 1/195)	1/75 (1/72 – 1/78)
MyCode – unrelated (n=109,730)	Number of individuals	2,489	728	1,751
	Prevalence	1/43 (1/41 – 1/44)	1/150 (1/140-1/162)	1/62 (1/59-1/65)

References

1. Stolz A, Ertych N, Bastians H: Tumor suppressor CHK2: regulator of DNA damage response and mediator of chromosomal stability. Clin Cancer Res 17:401-5, 2011

2. Hanson H, Astiazaran-Symonds E, Amendola LM, et al: Management of individuals with germline pathogenic/likely pathogenic variants in CHEK2: A clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). Genet Med 25:100870, 2023

3. Dorling L, Carvalho S, Allen J, et al: Breast cancer risks associated with missense variants in breast cancer susceptibility genes. Genome Med 14:51, 2022

4. Wilczewski CM, Obasohan J, Paschall JE, et al: Genotype first: Clinical genomics research through a reverse phenotyping approach. Am J Hum Genet 110:3-12, 2023

5. Bycroft C, Freeman C, Petkova D, et al: The UK Biobank resource with deep phenotyping and genomic data. Nature 562:203-209, 2018

6. Conroy MC, Lacey B, Besevic J, et al: UK Biobank: a globally important resource for cancer research. Br J Cancer 128:519-527, 2023

7. Carey DJ, Fetterolf SN, Davis FD, et al: The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med 18:906-13, 2016

8. Dewey FE, Murray MF, Overton JD, et al: Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science 354, 2016

9. Cingolani P, Platts A, Wang le L, et al: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80-92, 2012

10. Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164, 2010

11. Landrum MJ, Lee JM, Benson M, et al: ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46:D1062-D1067, 2018

12. Li Q, Wang K: InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet 100:267-280, 2017

13. Richards S, Aziz N, Bale S, et al: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405-24, 2015

14. Kim J, Naqvi AS, Corbett RJ, et al: AutoGVP: a dockerized workflow integrating ClinVar and InterVar germline sequence variant classification. Bioinformatics, 2024

15. Chow S-C, Chow S-C, Shao J, et al: Sample size calculations in clinical research (ed 2nd). Boca Raton, Chapman & Hall/CRC, 2008

16. Zhou W, Bi W, Zhao Z, et al: SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. Nat Genet 54:1466-1469, 2022

17. Meijers-Heijboer H, Wijnen J, Vasen H, et al: The CHEK2 1100delC mutation identifies families with a hereditary breast and colorectal cancer phenotype. Am J Hum Genet 72:1308-14, 2003

18. Bychkovsky BL, Agaoglu NB, Horton C, et al: Differences in Cancer Phenotypes Among Frequent CHEK2 Variants and Implications for Clinical Care-Checking CHEK2. JAMA Oncol 8:1598-1606, 2022

19. Weischer M, Bojesen SE, Tybjaerg-Hansen A, et al: Increased risk of breast cancer associated with CHEK2*1100delC. J Clin Oncol 25:57-63, 2007

20. Carlo MI, Mukherjee S, Mandelker D, et al: Prevalence of Germline Mutations in Cancer Susceptibility Genes in Patients With Advanced Renal Cell Carcinoma. JAMA Oncol 4:1228-1235, 2018

21. Han SH, Camp SY, Chu H, et al: Integrative Analysis of Germline Rare Variants in Clear and Non-clear Cell Renal Cell Carcinoma. Eur Urol Open Sci 62:107-122, 2024

22. Näslund-Koch C, Nordestgaard BG, Bojesen SE: Increased Risk for Other Cancers in Addition to Breast Cancer for CHEK2*1100delC Heterozygotes Estimated From the Copenhagen General Population Study. J Clin Oncol 34:1208-16, 2016

23. Pemov A, Wegman-Ostrosky T, Kim J, et al: Identification of Genetic Risk Factors for Familial Urinary Bladder Cancer: An Exome Sequencing Study. JCO Precis Oncol 5, 2021

24. Mian A, Wei J, Shi Z, et al: Systematic review of reported association studies of monogenic genes and bladder cancer risk and confirmation analysis in a large population cohort. BJUI Compass 4:156-163, 2023

25. Yang Y, Zhang G, Hu C, et al: The germline mutational landscape of genitourinary cancers and its indication for prognosis and risk. BMC Urol 22:196, 2022

26. Rudd MF, Sellick GS, Webb EL, et al: Variants in the ATM-BRCA2-CHEK2 axis predispose to chronic lymphocytic leukemia. Blood 108:638-44, 2006

27. Stubbins RJ, Korotev S, Godley LA: Germline CHEK2 and ATM Variants in Myeloid and Other Hematopoietic Malignancies. Curr Hematol Malig Rep 17:94-104, 2022

28. Zeng C, Bastarache LA, Tao R, et al: Association of Pathogenic Variants in Hereditary Cancer Genes With Multiple Diseases. JAMA Oncol 8:835-844, 2022

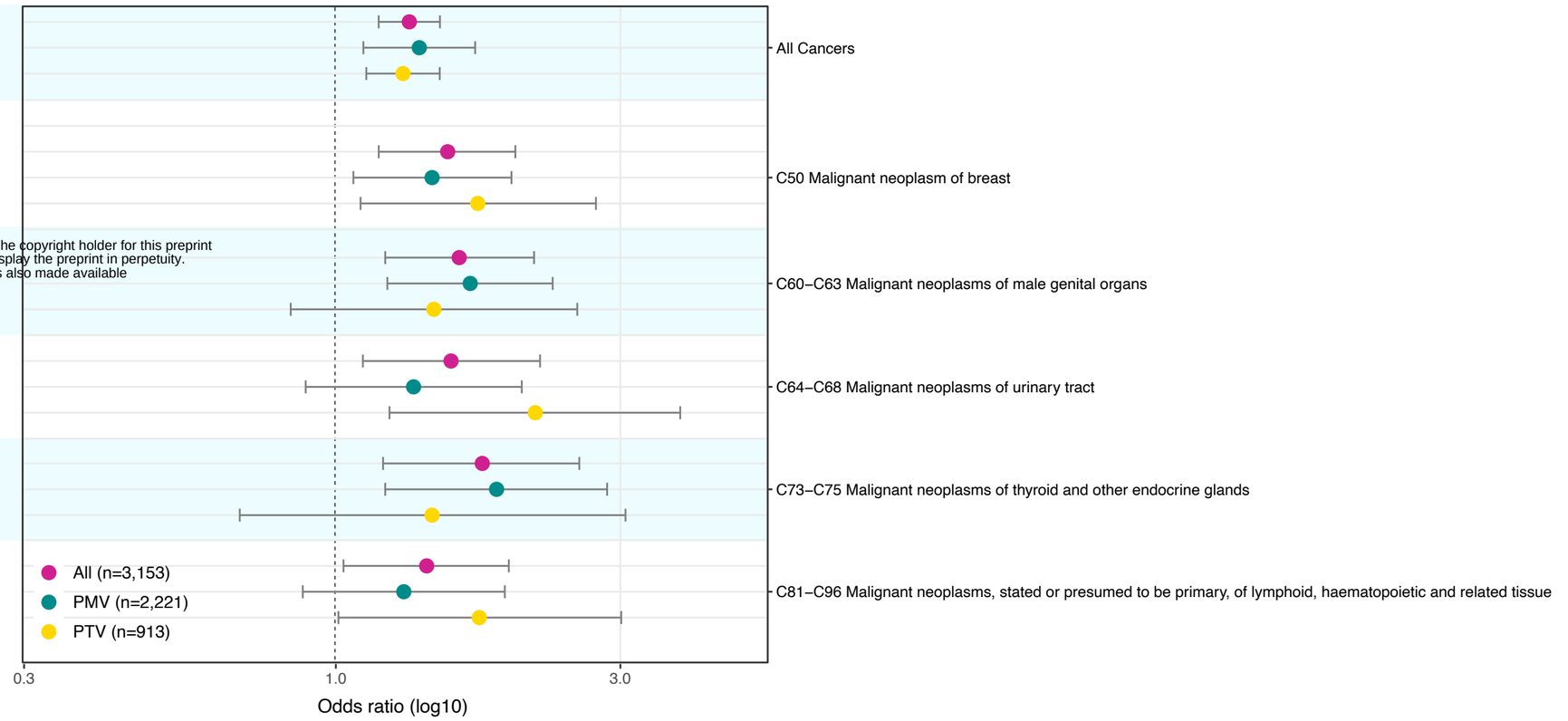
29. Kaczmarek-Ryś M, Ziemnicka K, Hryhorowicz ST, et al: The c.470 T > C CHEK2 missense variant increases the risk of differentiated thyroid carcinoma in the Great Poland population. Hered Cancer Clin Pract 13:8, 2015

30. Kim J, Haley J, Hatton JN, et al: A genome-first approach to characterize DICER1 pathogenic variant prevalence, penetrance and cancer, thyroid, and other phenotypes in 2 population-scale cohorts. Genet Med Open 2, 2024

31. Fry A, Littlejohns TJ, Sudlow C, et al: Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol 186:1026-1034, 2017

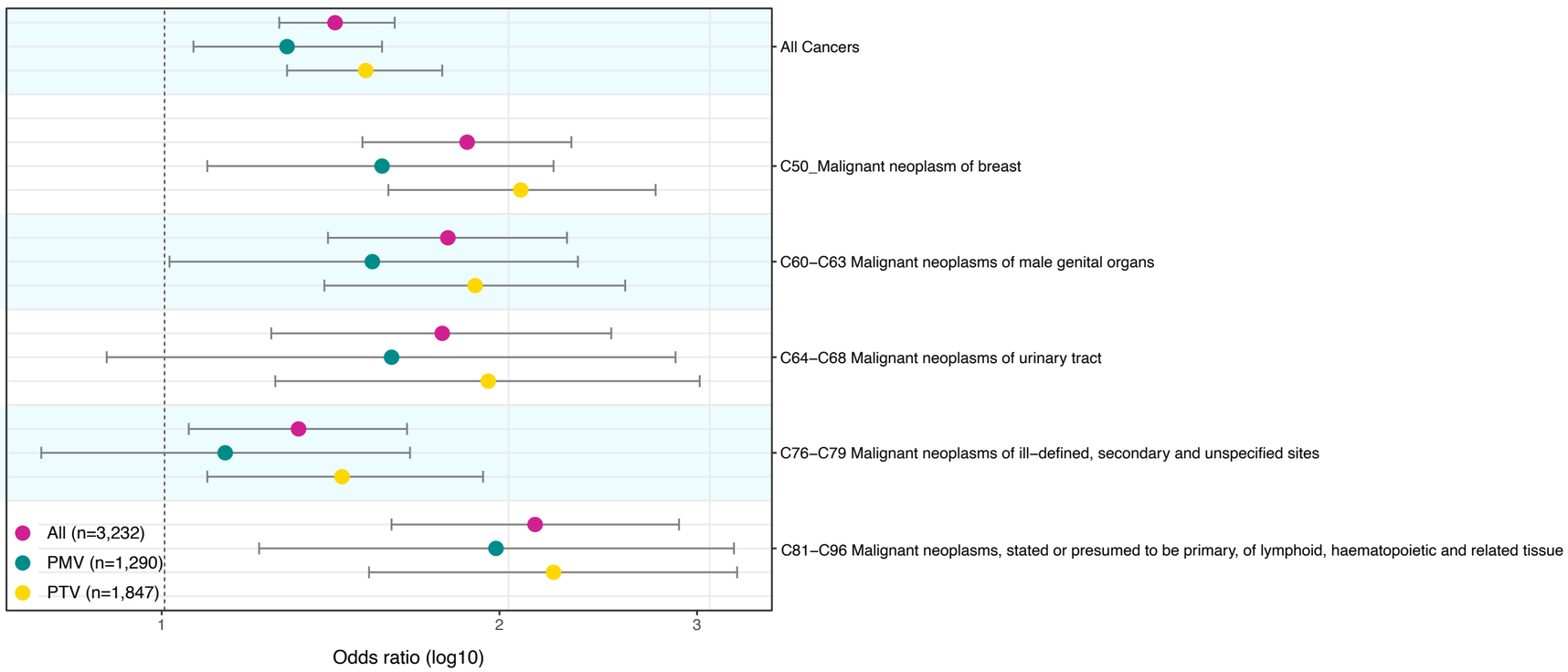
A.

Controls(%)	Heterozygotes(%)	OR [95% CI]	Adjusted SAIGE p–value
34,160(22.40)	864(27.40)	1.33[1.18–1.49]	2.15E–10
34,160(22.40)	609(27.40)	1.38[1.11–1.71]	2.84E–07
34,160(22.40)	247(27.1)	1.30[1.12–1.49]	1.09E–04
5,004(3.28)	156(4.95)	1.54[1.18–2.00]	2.45E–05
5,004(3.28)	105(4.73)	1.45[1.07–1.97]	1.05E–02
5,004(3.28)	49(5.37)	1.73[1.10–2.73]	0.01
4,443(7.49)	142(11.66)	1.61[1.21–2.15]	2.05E–04
4,443(7.49)	106(12.08)	1.68[1.22–2.30]	6.90E–04
4,443(7.49)	36(10.74)	1.46[0.84–2.54]	0.96
2,632(1.72)	84(2.66)	1.56[1.11–2.2]	3.76E–03
2,632(1.72)	53(2.39)	1.35[0.89–2.05]	0.7
2,632(1.72)	31(3.40)	2.16[1.23–3.78]	0.03
1,840(1.21)	66(2.09)	1.76[1.2–2.56]	9.11E–04
1,840(1.21)	49(2.21)	1.86[1.21–2.85]	1.11E–03
1,840(1.21)	16(1.75)	1.45[0.69–3.06]	1
3,107(2.04)	91(2.88)	1.42[1.03–1.95]	0.04
3,107(2.04)	60(2.70)	1.3[0.88–1.92]	1
3,107(2.04)	31(3.40)	1.74[1.01–3.01]	0.06



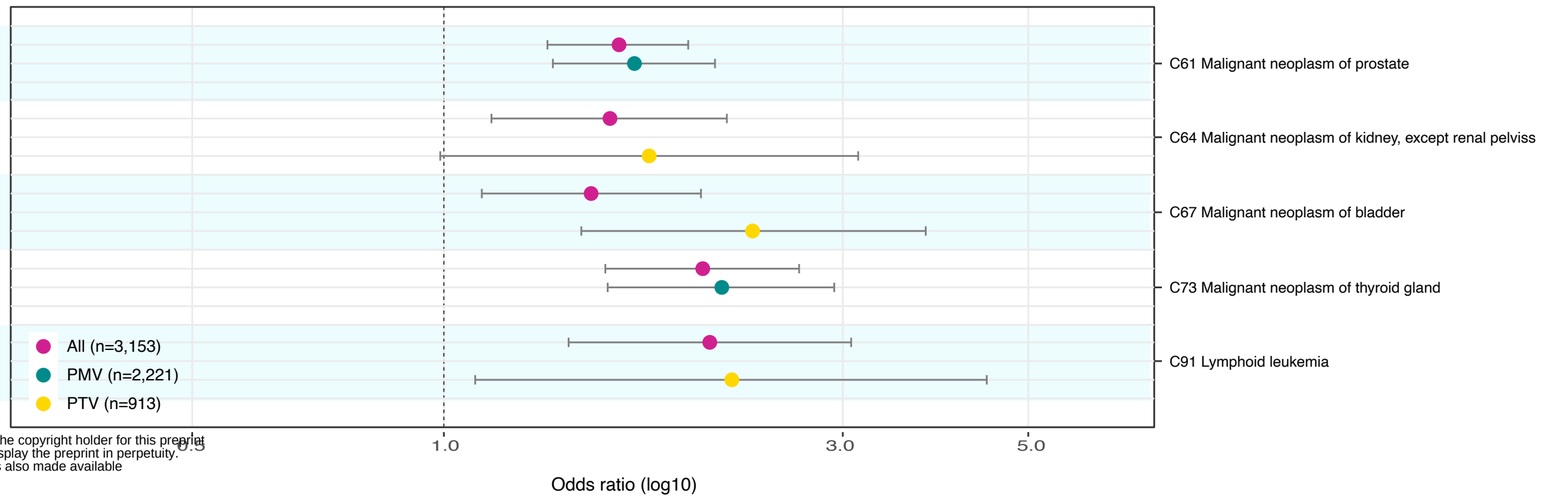
B.

Controls(%)	Heterozygotes(%)	OR [95% CI]	Adjusted SAIGE p–value
69,836(22.87)	934(28.90)	1.41[1.26–1.59]	6.05E–15
69,836(22.87)	344(26.67)	1.28[1.06–1.55]	2.19E–04
69,836(22.87)	562(30.23)	1.50[1.28–1.75]	9.56E–12
12,439(4.07)	227(7.02)	1.84[1.49–2.27]	2.47E–12
12,439(4.07)	80(6.20)	1.55[1.09–2.19]	1.41E–02
12,439(4.07)	139(7.53)	2.05[1.57–2.69]	4.49E–10
10,557(3.46)	179(5.53)	1.77[1.39–2.25]	3.64E–09
10,557(3.46)	58(4.50)	1.52[1.01–2.30]	0.14
10,557(3.46)	113(6.12)	1.87[1.38–2.53]	4.13E–07
4,351(1.43)	77(2.38)	1.75[1.24–2.46]	1.01E–04
4,351(1.43)	27(2.09)	1.58[0.89–2.80]	0.84
4,351(1.43)	49(2.65)	1.92[1.25–2.94]	2.68E–04
14,339(4.70)	192(5.94)	1.31[1.05–1.63]	4.68E–03
14,339(4.70)	66(5.12)	1.13[0.78–1.64]	1
14,339(4.70)	120(6.50)	1.43[1.09–1.90]	2.95E–03
5,030(1.65)	108(3.34)	2.11[1.58–2.82]	3.45E–09
5,030(1.65)	39(3.02)	1.95[1.21–3.15]	4.70E–03
5,030(1.65)	65(3.52)	2.19[1.51–3.17]	8.74E–07



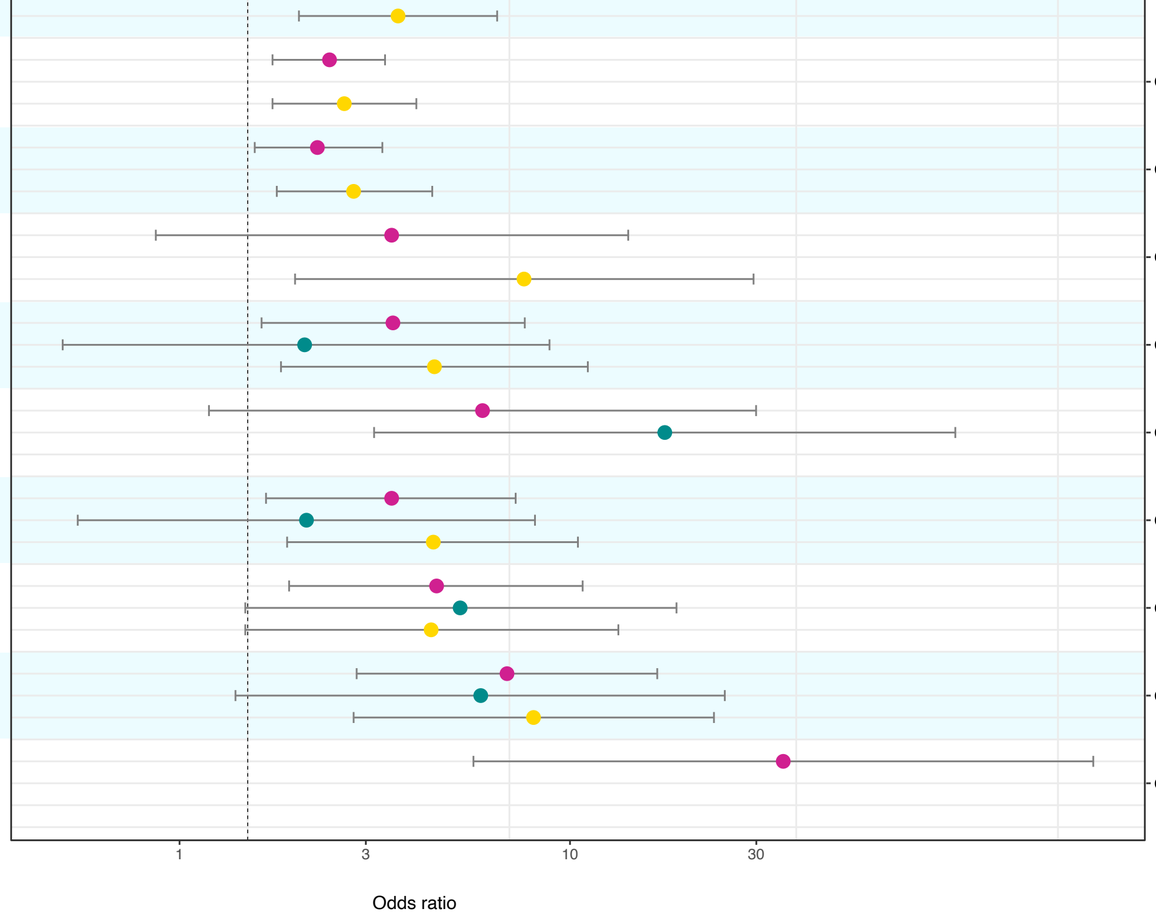
A.

Controls(%)	Heterozygotes(%)	OR [95% CI]	Adjusted SAIGE p–value	MyCode
4222(7.12)	135(11.09)	1.62[1.27–2.07]	1.04E–04	C61 Malignant neoplasm of prostate
4222(7.12)	101(11.51)	1.69[1.27–2.24]	4.03E–04	
4222(7.12)	34(10.14)	NA	NA	
1202(0.79)	39(1.24)	1.58[1.03–2.41]	0.02	C64 Malignant neoplasm of kidney, except renal pelviss
1202(0.79)	27(1.22)	NA	NA	
1202(0.79)	12(1.31)	1.76[0.83–3.75]	0.28	
1477(0.96)	46(1.46)	1.50[1.01–2.23]	0.1	C67 Malignant neoplasm of bladder
1477(0.96)	27(1.22)	NA.	NA	
1477(0.96)	19(2.08)	2.34[1.254.38]	3.51E–03	
1370(0.90)	57(1.81)	2.04[1.48–2.82]	1.82E–05	C73 Malignant neoplasm of thyroid gland
1370(0.90)	42(1.89)	2.15[1.47–3.14]	6.60E–05	
1370(0.90)	14(1.53)	NA	NA	
627(0.41)	27(0.86)	2.08[1.17–3.69]	7.61E–03	C91 Lymphoid leukemia
627(0.41)	19(0.86)	NA	NA	
627(0.41)	8(0.88)	2.21[0.78–6.22]	0.53	

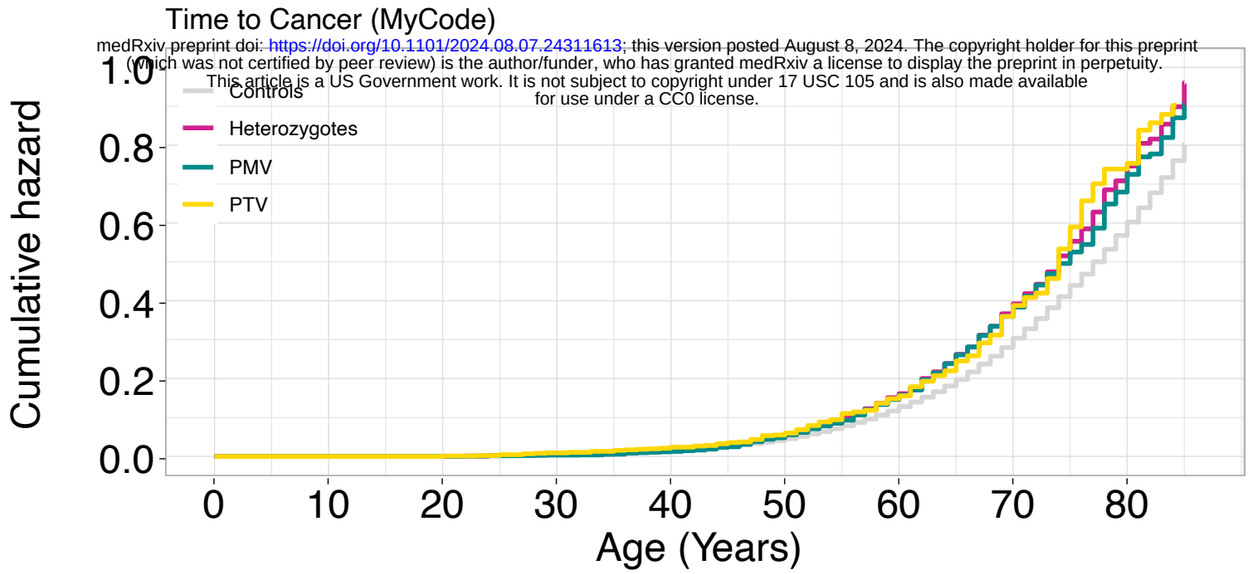


B.

Controls(%)	Heterozygotes(%)	OR [95% CI]	Adjusted SAIGE p–value	UK Biobank
9,978(3.27)	170(5.27)	1.78[1.48–2.16]	6.68E–10	C61 Malignant neoplasm of prostate
9,978(3.27)	55(4.26)	NA	NA	
9,978(3.27)	107(5.79)	1.87[1.47–2.38]	1.27E–07	
1,635(0.54)	31(0.96)	1.84[1.22–2.77]	2.49E–03	C64 Malignant neoplasm of kidney, except renal pelvis
1,635(0.54)	11(0.85)	NA	NA.	
1,635(0.54)	20(1.08)	2.05[1.23–3.40]	4.28E–03	
2,707(0.89)	45(1.40)	1.64[1.17–2.31]	2.66E–03	C67 Malignant neoplasm of bladder
2,707(0.89)	15(1.16)	NA	NA	
2,707(0.89)	30(1.62)	1.88[1.24–2.85]	2.32E–03	
7,493(2.45)	109(3.38)	1.41[1.11–1.78]	9.77E–04	C77 Secondary and unspecified malignant neoplasm of lymph nodes
7,493(2.45)	40(3.10)	NA	NA	
7,493(2.45)	66(3.57)	1.50[1.11–2.03]	4.23E–03	
6,072(2.00)	84(2.61)	1.34[1.03–1.76]	0.01	C79 Secondary malignant neoplasm of other sites
6,072(2.00)	27(2.09)	NA	NA	
6,072(1.99)	56(3.03)	1.56[1.13–2.17]	4.40E–03	
414(0.14)	8(0.25)	1.83[0.68–4.94]	0.02	C81 Hodgkin's disease
414(0.14)	0(0)	NA	NA	
414(0.14)	8(0.43)	3.19[1.22–8.36]	8.40E–03	
1,363(0.45)	26(0.81)	1.84[1.06–3.20]	0.03	C83 Diffuse non–Hodgkin's lymphoma
1,363(0.45)	7(0.54)	1.27[0.46–3.55]	1	
1,362(0.45)	18(0.97)	2.19[1.15–4.17]	0.01	
216(0.07)	6(0.19)	2.68[0.85–8.45]	0.24	C84 Peripheral and cutaneous T–cell lymphomas
216(0.07)	5(0.39)	5.76[1.70–19.50]	0.02	
216(0.07)	1(0.05)	NA	NA	
1,533(0.50)	29(0.90)	1.83[1.08–3.08]	0.02	C85 Other and unspecified types of non–Hodgkin's lymphoma
1,533(0.50)	8(0.62)	1.28[0.49–3.34]	0.94	
1,533(0.50)	20(1.08)	2.18[1.18–4.00]	0.02	
921(0.30)	21(0.65)	2.21[1.19–4.08]	0.01	C91 Lymphoid leukaemia
921(0.30)	9(0.70)	2.44[0.99–6.05]	0.14	
921(0.30)	12(0.65)	2.16[0.99–4.74]	0.08	
650(0.21)	20(0.62)	2.97[1.58–5.58]	2.17E–04	C92 Myeloid leukaemia
650(0.21)	7(0.54)	2.66[0.95–7.41]	0.13	
650(0.21)	13(0.70)	3.32[1.56–7.08]	1.22E–03	
50(0.02)	5(0.16)	9.47[2.58–34.80]	1.91E–03	C96 Other and unspecified malignant neoplasms of lymphoid, haematopoietic and related tissue
50(0.02)	3(0.23)	NA	NA	
50(0.02)	2(0.11)	NA	NA	



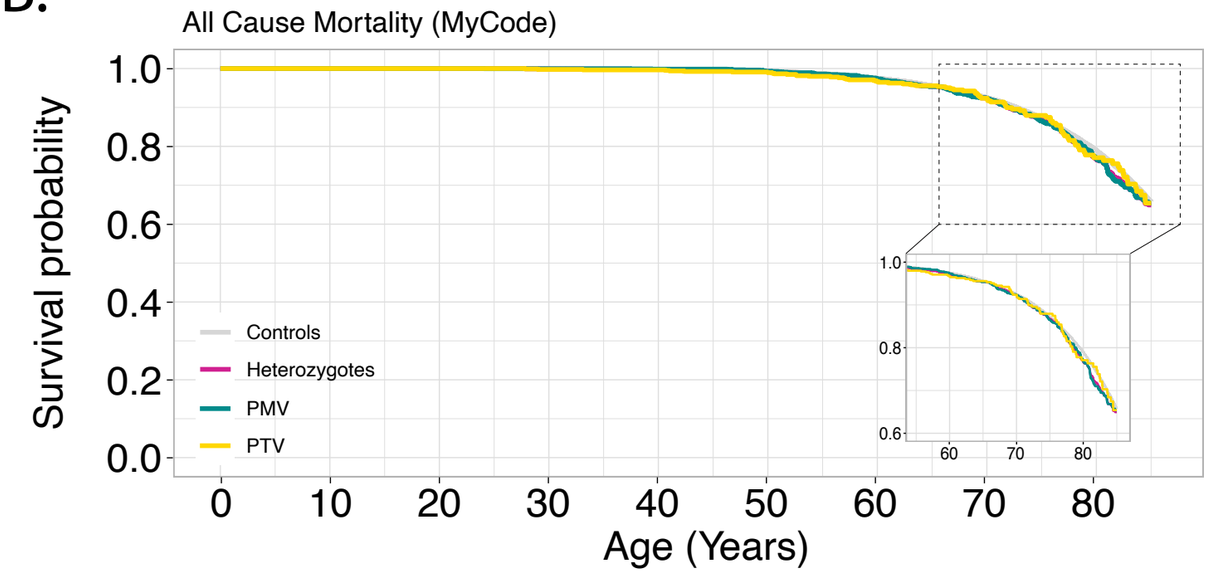
A.



Number at risk

Controls	148049	148047	147563	138641	121024	100185	71541	38520	12971
Heterozygotes	3021	3021	3015	2837	2497	2073	1446	759	241
PMV	2141	2141	2135	2014	1772	1486	1050	551	180
PTV	881	881	881	824	726	592	403	218	69

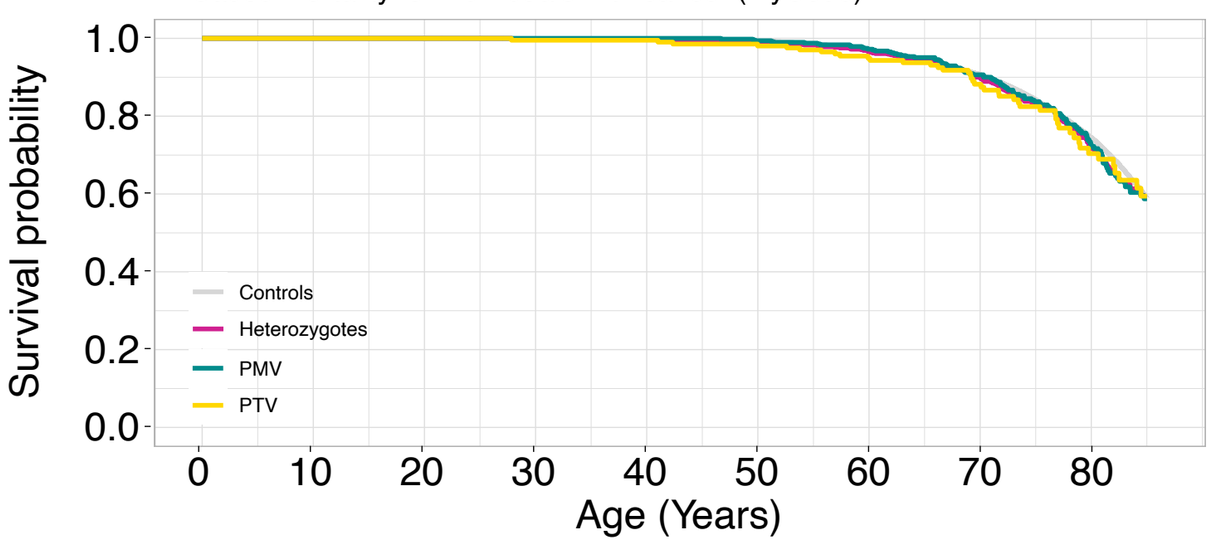
B.



Number at risk

Controls	151975	151973	151534	143014	126032	106622	80626	48585	19658
Heterozygotes	3133	3133	3128	2958	2631	2257	1702	1041	402
PMV	2221	2221	2216	2099	1866	1616	1233	760	299
PTV	913	913	913	860	766	643	472	288	110

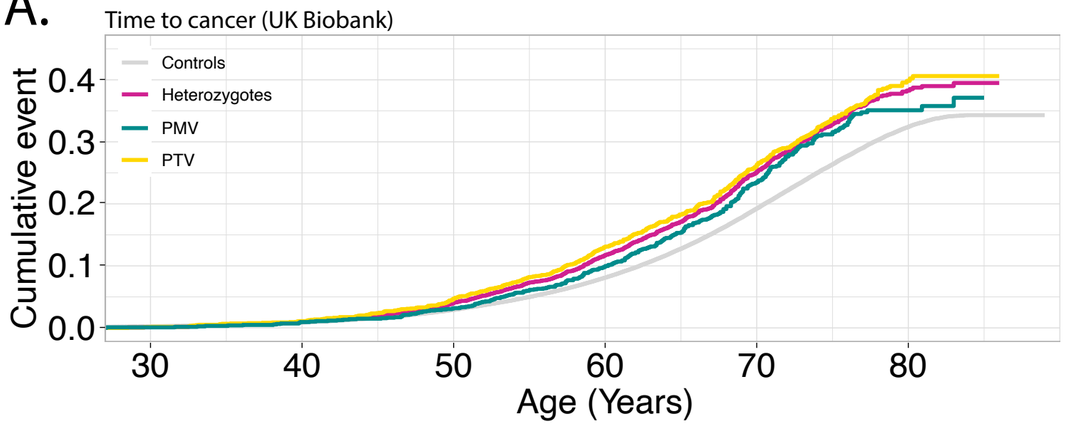
C.



Number at risk

Controls	29280	29280	29279	29174	28633	27276	23886	17039	7755
Heterozygotes	732	732	732	726	716	687	599	419	175
PMV	512	512	512	510	505	485	423	298	125
PTV	212	212	212	208	203	195	171	118	49

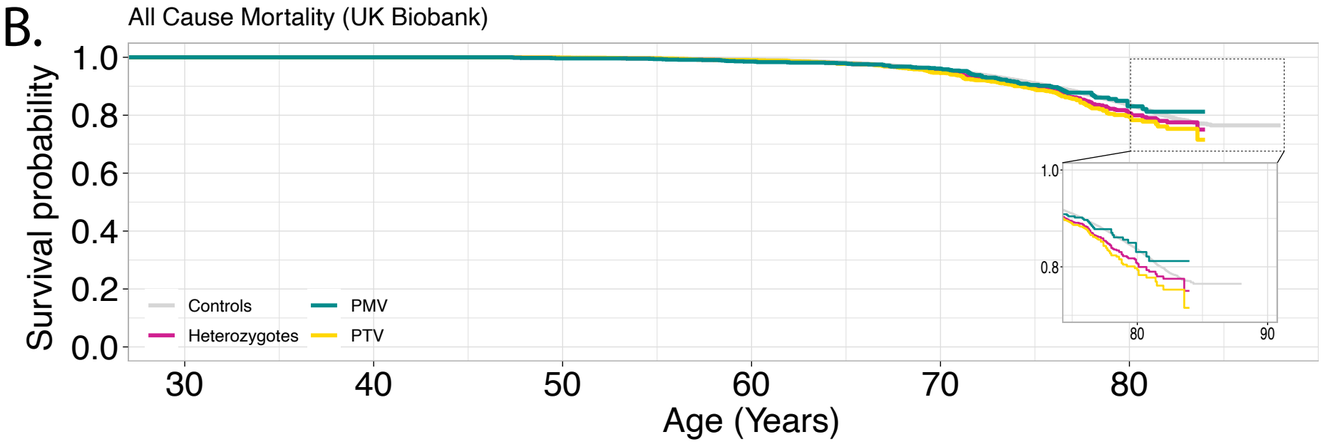
A.



Number at risk

Controls	284053	302927	296571	251558	151143	37197
Heterozygotes	3016	3192	3096	2554	1480	357
PMV	1275	1278	1250	1040	589	132
PTV	1737	1839	1773	1453	855	218

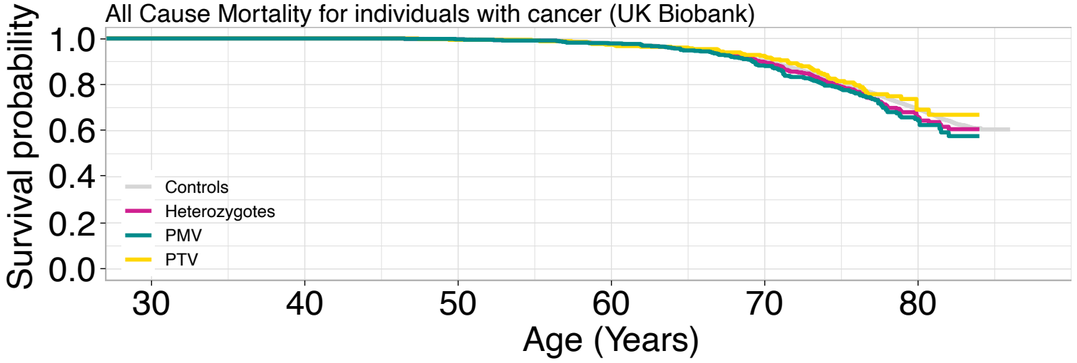
B.



Number at risk

Controls	305311	305311	304854	263934	170315	36963
Heterozygotes	3224	3224	3216	2767	1779	356
PMV	1290	1290	1285	1099	706	131
PTV	1859	1859	1856	1602	1032	215

C.



Number at risk

Controls	69836	69836	69669	65532	49440	12539
Heterozygotes	934	934	930	869	638	140
PMV	562	562	560	521	377	91
PTV	344	344	342	321	242	45