A unified framework for estimating country-specific cumulative incidence for 18 diseases stratified by polygenic risk

Supplementary Methods

Study Specific Quality Control

UK Biobank

Registry data
The relevant columns used to define the phenotypes were:

- Cause of Death Primary (Column ID: 40001)
- Cause of Death Secondary (Column ID: 40002)
- Summary ICD10 Diagnoses (Column ID: 41270)
- Summary ICD10 Diagnoses Date (Column ID: 41280)
- Summary ICD9 Diagnoses (Column ID: 41271)
- Summary ICD9 Diagnoses Date (Column ID: 41281)

Such data are taken from the hospital episode statistics which relate to hospital inpatient data. For more information, please see here. Registry coverage depends on the country with follow-up beginning in 1997, 1998 and 1981 for England, Wales and Scotland respectively. End of follow-up was stated as 31st January 2021.

Genotyping and quality control
Two arrays were used to genotype UK Biobank participants. The UK Biobank Lung Exome Variant Evaluation (UKBiLEVE) Axiom array was used to genotype 49,950 participants. The remaining 438,427 participants were genotypes using the Applied Biosystems UK Biobank Axiom Array.

Principal Component Analysis (PCA) was performed on the genetic data and centralised quality control (QC) on variants was performed on individuals identified to belong to the largest cluster (N=463,844) according to Aberrant - an unsupervised clustering algorithm[1]. Variants were assessed for evidence of allele frequency variation across batch, plate, sex or array and that genotypes were largely consistent with Hardy-Weinberg Equilibrium expectations (all p-value thresholds < 10^{-12}). If a variant failed one or more tests within a given batch it was set to missing. Previous research[2] provides more detailed information on testing.

Imputation
For 487,442 individuals, imputation was performed using the IMPUTE4 [3] software. Genetic variation from the Haplotype Reference Consortium (HRC) [4] and merged UK10K+1000 Genomes were used as a reference panel[5]. Single Nucleotide Polymorphisms (SNPs) were only included in the final imputation if they were present in both reference panels, giving a total of 96,959,328 SNPs.

Ancestry assignment

Ancestry assignment uses methodology and scripts from GenoPred (https://opain.github.io/GenoPred/DiverseAncestry.html). Individuals were stratified into one of five super populations African (AFR), American (AMR), South Asian (SAS), East Asian (EAS) and European (EUR). The 1000 Genomes data [6] acted as a reference given the individuals are known to belong to one of the 5 super populations. Only unambiguous SNPs also present in both the HapMap3 consortium [7] and the imputed UK Biobank data were retained for PCA. SNPs within both the reference (1000 Genomes) and target (UK Biobank) samples underwent quality control such that the minor allele frequency (MAF) > 5%, variant missingness > 2% and Hardy-Weinberg Equilibrium p-value > $1e^{-6}$. 467,970 autosomal SNPs remained following QC and were in the intersection of the reference and target samples. Regions with long range linkage disequilibrium were excluded and independent SNPs (SNPs greater than 1000kb apart and $r^2 < 0.2$) retained. PCA was then performed in the reference sample using PLINK v2 [8] and a multinomial elastic-net regression was trained using 5-fold cross validation, super population as the outcome and the first 10 PCs as covariates. PCs from the target sample were then projected into the reference space and prediction on super population made. Classifications were made according to the super population with the greatest probability. To be classified the max probability must be over 0.5, otherwise it was set to missing.

PCA was performed using a random subset of 1000 individuals per super population and PC's from the rest of the super population sample projected onto this space. Distances from the centroid were calculated and outliers removed. Outliers were classified as having a distance > 75 percentile + 30*Interquartile Range. Following within-ancestry QC, 8,381, 1,063, 2,393, 447,332 and 9,435 individuals were allocated to AFR, AMR, EAS, EUR and SAS super populations respectively.

FinnGen

Registry data

Phenotype data within FinnGen is constructed from the collection of nationwide electronic health registers. This gives a comprehensive coverage of almost all of a patient's interactions with the health service including hospitalizations, medications, procedures and deaths. The 18 different registers used by the project are listed below in order of their follow-up times:

- Finnish Cancer Registry - From 1953
- Register of Congenital Malformations - From 1963
- Reimbursement - From 1964
- Population Register - From 1964
- Finnish Registry for Kidney Diseases - From 1964
- Causes of Death - From 1969
- Care Register for Health Care Inpatient Visits, HILMO - From 1969
- Socio-economic data - From 1970
- The Finnish Registry of Visual Impairment - From 1983
- Medical Birth Register - From 1987
- Finnish National Infectious Disease Register - From 1989
- Cervical Cancer Screening - From 1991
- Breast Cancer Screening - From 1992
- Drug Purchases - From 1995
- The Care Register for Social Welfare - From 1995
- Care Register for Health Care, specialist outpatient visits, HILMO - From 1998
- Register of Primary Health Care Visits, Avohilmo - From 2011

- The Finnish Vaccination Register - From 2011

Note: while primary health care visits are included within FinnGen, by default these cases are excluded from the endpoints. As such, we only consider secondary care data for our disease endpoints.

Genotyping and quality control
FinnGen consists of prospectively recruited samples and a series of legacy cohorts with genotypes already available [9]. Prospective samples were genotyped using the ThermoFisher Axiom custom array which tags a total of 655,973 variants. Genotype calling was performed using the Array Power Tools software. Legacy cohorts were genotyped using various Illumina arrays and genotype calling was performed using either GenCall or zCall algorithms.

For both prospective and legacy cohorts the following quality control metrics were used.

Samples were removed if:
- Pihat was > 0.9 and the samples were not monozygotic or replicates
- There was a discrepancy between reported sex and genetically determined sex (F-value ≤ 0.3 for females and ≥ 0.8 for males)
- Missingness was ≥ 5%
- Heterozygosity was ±4 standard deviations from the population average
- Pihat was > 0.1 with 14 or more samples
- Samples were ±4 standard deviations away from the population average according to the first two genetic principal components.

Samples were tagged should there be evidence of a mendelian error or contain replicate samples with over 50,000 discrepancies.

Variants were removed if:
- The variant failed the Hardy-Weinberg Equilibrium test (p-value < 10^{-6})
- The variant had a call rate < 98%

Imputation
Pre-phasing was performed using Eagle 2.3.5[10] and samples were imputed using the SiSu v3 imputation reference panel. This reference panel is specific to the Finnish population, containing high-coverage (25-30x) whole-genome sequencing data from 3,775 Finns and 16,962,023 variants with minor allele count ≥ 3. After imputation, 16,387,711 variants were imputed with high quality (INFO > 0.6)

Ancestry assignment
Firstly, the FinnGen samples were combined with the 1000 genomes phase 3 dataset[6]. Genetic principal components were calculated using a subset of 49,451 pruned SNPs. Aberrant[1] was used to identify and remove samples that deviated from the main cluster. A probability of belonging to either a North-Western European or Finnish population was calculated by firstly performing PCA with individuals belonging to these ancestries from 1000 genomes data. FinnGen samples were then projected onto this PCA space and Mahalanobis distances calculated for each sample against each of the two ancestries. Samples were retained if there was ≥ 95% probability of belonging to the Finnish ancestry cluster.

Trøndelag Health Study

Registry data

The periodic population-based health survey design includes three recruitment waves—HUNT1 (1984-1986), HUNT2 (1995-1997), and HUNT3 (2006-2008)—concentrated in the North-Trøndelag area, where all adults > 20 years of age were invited to participate. Electronic health records from the Trøndelag county hospitals (Nord-Trøndelag Hospital Trust, including St. Olavs, Namsos, and Levanger Hospitals) hold International Classification of Diseases and Related Health Problems (ICD) codes back to 1987 and were last accessed August 8, 2021. There is likely under-ascertainment of less-serious common conditions with only hospital records. Main and secondary diagnoses were used. Follow-up time is calculated from age of first enrollment in HUNT to the last hospital record accession.

Genotyping and quality control

DNA from 71,860 HUNT samples was genotyped using one of three different Illumina HumanCoreExome arrays (HumanCoreExome12 v1.0, HumanCoreExome12 v1.1 and UM HUNT Biobank v1.0). These chips included custom content to directly genotyped missense and loss of function variants and lipid associated variants from low pass sequencing, among other potentially actionable protein altering variants. Samples that failed to reach a 99% call rate, had contamination > 2.5% as estimated with BAF Regress [11], large chromosomal copy number variants, lower call rate of a technical duplicate pair and twins, gonosomal constellations other than XX and XY, or whose inferred sex contradicted the reported gender, were excluded. Samples that passed quality control were analysed in a second round of genotype calling following the Genome Studio quality control protocol described elsewhere [12]. Genomic position, strand orientation and the reference allele of genotyped variants were determined by aligning their probe sequences against the human genome (Genome Reference Consortium Human genome build 37 and revised Cambridge Reference Sequence of the human mitochondrial DNA; http://genome.ucsc.edu) using BLAT. Variants were excluded if their probe sequences could not be perfectly mapped to the reference genome, cluster separation was < 0.3, Gentrain score was < 0.15, showed deviations from Hardy Weinberg equilibrium in unrelated samples of European ancestry with p-value < 0.0001), their call rate was < 99%, or another assay with higher call rate genotyped the same variant.

Imputation

Imputation was performed on the 69,716 samples of recent European ancestry using Minimac3 (v2.0.1, http://genome.sph.umich.edu/wiki/Minimac3) [13] with default settings (2.5 Mb reference based chunking with 500kb windows) and a customized Haplotype Reference consortium release 1.1 (HRC v1.1) for autosomal variants and HRC v1.1 for chromosome X variants[4]. The customized reference panel represented the merged panel of two reciprocally imputed reference panels: 1) 2,201 low-coverage whole-genome sequences samples from the HUNT study and 2) HRC v1.1 with 1,023 HUNT WGS samples removed before merging. We excluded imputed variants with Rsq < 0.3 resulting in over 24.9 million well-imputed variants.

Ancestry assignment

Ancestry of all samples was inferred by projecting all genotyped samples into the space of the principal components of the Human Genome Diversity Project (HGDP) reference panel (938 unrelated individuals; downloaded from http://csg.sph.umich.edu/chaolong/LASER/) [14], using PLINK v1.90 (8). Recent European ancestry was defined as samples that fell into an ellipsoid spanning exclusively European populations of the HGDP panel. The different arrays were harmonized by reducing to a set of overlapping variants and excluding variants that showed frequency differences > 15% between data sets, or that were monomorphic in one and had MAF > 1% in another data set. The resulting genotype data were phased using Eagle2 v2.3 [10].

Estonian Biobank

Registry data

Phenotype data within EstBB is put together from the collection of electronic health registers, including from two largest hospitals in Estonia. We include both primary and secondary care data as well as self-reported diagnoses. Estonia has a solidary health insurance system and national public health insurance covers ~94% of the population (https://eurohealthobservatory.who.int/countries/estonia). Causes of death and Cancer registry record all cases despite the health insurance status in Estonia. Following registries were included in phenotype definition process:

- Causes of Death Registry- diagnoses from 2003 until 2020
- National Cancer Registry- diagnoses from 1955 until 2017
- Estonian Health Insurance Fund - From 2001 until 2020
- The North Estonia Medical Centre from 1993 until 2017
- Tartu University hospital from 2006 until 2017
- E-Health system from 1998 to 2020

Self-reported diagnoses' dates ranged from 1920 - 2018. For the follow-up time calculation in Table 1, we used as baseline the start of National Health Insurance Fund data from 2003 (2001-2002 no state-wide coverage) until the end of last linking in 2020, so the follow-up IQR is 0.

Genotyping and quality control

Estonian BioBank (EstBB) samples were genotyped with 4 sub-versions of Infinium Global Screening Array-24. Samples with less than 95% call-rate were excluded. Sample sex recorded in the EstBB database was compared with genetic sex. Samples with sex mismatch were further inspected for sex chromosome abnormalities (X0, XXY, etc.), and samples with confirmed database vs genetic sex mismatch were excluded. In total, 202 910 individuals passed sample quality control. SNP quality control was performed by excluding: (a) all SNPs with less than 95% call-rate, (b) SNPs showing more than 5% AF difference from the AF mean estimated using all genotyping batches with more than 10 000 samples per batch, (c) SNPs with Illumina GenTrain score < 0.6 or cluster separation score < 0.4 in any genotyping batch, (d) autosomal SNPs with HWE exact test p-value < 1e-4. In total, approximately 328K autosomal and X-chromosome SNPs with MAF > 1% passed quality control and were used in the imputation. All the variants were processed on the human genome assembly GRCh37.

Imputation

Imputation was performed using local Estonian imputation reference panel made of 2056 WGS samples. Genotypes were pre-phased with Eagle v2.4.1 (10) and imputed with Beagle 5.1 using default parameters[15]. Multiallelic positions were excluded from imputation output. In total, 39 546 641 variants were used in the study.

Ancestry assignment

EstBB samples were combined with the 1000 genomes phase 3 dataset for ancestry analysis[6]. Genetic principal components were calculated using a subset of quality controlled and pruned genotyped SNPs. This was further used to identify and remove samples that deviated from the main cluster via visual inspection. In total, 481 non-european ancestry individuals based on principal components were excluded from the analysis.

Mass General Brigham Biobank

Registry data

Patients and employers of multiple health centers at Mass General Brigham (MGB) in Eastern Massachusetts are enrolled in the MGB Biobank[16]. The MGB Biobank

was founded in 2008 and the research protocol was approved by the Human Research Committee of MGB. The EHR data were retrieved from the MGB Patient Data Registry (RPDR). A biobank portal which is an i2b2-based data repository linking disparate and high-dimensional patient data was implemented . The weekly updated repository integrates Information from primary and curated data resources. Data stored on the observational medical outcomes partnership (OMOP) can also be queried through i2b2. The start of follow-up was defined as the first diagnosed ICD code across all EHG data. The end of follow-up was defined as death date for deceased people and most recent visit date for others. The ICD codes were available in the MGBB since its establishment in 2008.

Genotyping and quality control
To date, ~65,000 individuals with informed consent have generated and provided genomic information. In this study, we were focused on those individuals genotyped on the Illumina Global Screening Array. We retained genotyped SNPs that: i) minor allele count >= 2, ii) missingness <= 2% iii) Hardy-Weinberg equilibrium *p*-value >= $1e^{-6}$, and iv) concordant allele frequency with gnomAD (chi-squared value >= 300). For individual-level quality controls, we removed those individuals that i) heterozygosity > 3 standard deviation of population mean, ii) missingness > 1% and ii) discordant sex between self-reported and genetic inferred sex. Finally, 563,449 genotyped variants for 52,459 individuals were subsequently imputed.

Imputation
The genotypes after quality controls were imputed at the Michigan imputation server using the TOPMed r^2 imputation panel [17] using Minimac [18]. Eagle v2.4 was used for haplotype phasing[10]. Variants with imputation INFO scores > 0.3 were further retained for follow-up analyses.

Ancestry assignment
We projected all individuals onto the first 6 PCs based on 168,898 variants in the combined reference dataset from 1000 Genomes Project Phase 3 and Human Genome Diversity Project. We applied a random forest classifier and assigned ancestry to 6 continental groups (including European, Central and South-Asian, East-Asian, African, Middle-Eastern and American) if the probability was larger than 0.8.

Genomics England

Registry data
Available clinical data in the Genomics England research environment are divided into primary, sourced from the Genomic Medicine Centres for all participants upon enrollment in the program, and secondary clinical data come from third parties such as Public Health England or NHSD which complement the primary clinical data with additional information. The following table was included in the phenotype definition process:

Hospital Episode Statistics admitted patient care (hes_apc) - contains historic records of admissions into secondary care of Genomics England main programme participants. The period covered by the registry is from September 13, 1992, to January 31, 2022 (Genomics England version 15). The HES records are based on the ICD-10 coding system, which has been in use since April 1995. For more information please see: https://re-docs.genomicsengland.co.uk/release15/.The beginning of follow-up was defined as the date for registry linkage. If an individual was born after that date, the date of birth is the beginning of follow-up time.

Genotyping and quality control
Genome sequencing was performed in DNA samples from 78,195 individuals using Illumina HiSeq X systems (150 bp paired-end format). Reads were aligned using the iSAAC Aligner (version 03.16.02.19) and small variants were called using Starling Small Variant Caller (version 2.4.7). Samples were aligned to the Homo Sapiens NCBI GRCh38 assembly with decoys.

Aggregation of single-sample gVCFs was performed using the Illumina software gVCF genotyper (version 2019). Variant normalisation and decomposition were implemented by vt (version 0.57721). Genomic annotation and calculation of allele statistics were performed using Ensembl VEP and bcftools respectively. The multi-sample VCF dataset (aggV2) was then split into 1,371 roughly equal chunks to allow faster processing. Only variants that passed all provided site quality control criteria were processed.

Imputation
The WGS genotypes (~722M variants) were filtered to a variant base list used for PGS model generation, which includes 18,421,839 variants. (For further information on how the variant list was derived see: https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=72351761)
Genotypes were phased and imputed using the 1000G reference panel (v5a) which was lifted-over from GRCh37 to GRCh38 using cross-map.

Ancestry assignment
The genetic ancestry of the patients was estimated using a random forest classifier and data from 1000 genomes project phase 3 (1KGP3) dataset. Firstly, all unrelated samples from the 1KGP3 were selected and 188,382 HQ SNPs were subsetted. After filtering for MAF > 0.05 in 1KGP3 (and GE data), the first 20 PCs were calculated using GCTA and the aggV2 data were projected onto the 1KGP3 PC loadings. The random forest model to predict ancestries was trained based on:

- First 8 1KGP3 PCs
- set Ntrees = 400
- Train and predict on 1KGP3 Admixed American, African, East Asian, European, and South Asian super-populations.

Individuals were assigned for any one ancestry with a probability of > 0.8.

Generation Scotland

Registry data
Disease outcomes were ascertained through linkage to primary (GP) and secondary (hospital) healthcare records. Individuals were subsetted to those registered at a GP that consented to sharing of primary records. GP records consisted of Read2 codes, which were mapped to ICD-10. Hospital data were obtained from Scottish Morbidity Records (SMR) where disease outcomes were coded using ICD-9 (pre March 1997) or ICD-10 (post March 1997). Start of followup was considered to be the latest date between Date of birth or March 1980, the date of GP linkage. End of followup was considered to be October 2020 (the date to which GP data is available), date of death, or date of disease onset if after October 2020 (hospital data is available until March 2022).

Genotyping and quality control

Generation Scotland (GS) consists of ~24,000 individuals from across Scotland aged between 18-99 years. Phenotypic data were obtained at baseline along with whole blood samples for DNA quantification.

Genotype data was assayed for 20,195 participants in two batches with 9,863 participants in the first batch and the remainder in the second. The genotyping was performed using the Illumina HumanOmniExpressExome-8 v1.0 BeadChip and the Illumina HumanOmniExpressExome-8 v1.2 BeadChip, respectively. Individuals or SNPs with a low call rate (<98%) and SNPs with Hardy-Weinberg p-value<$1x10^{-6}$ were removed. Mendelian errors were removed by setting the individual-level genotypes at erroneous SNPs to missing.

Imputation
Genotyped data were imputed using the HRC panel v1.1[4]. Autosomal haplotypes were checked to ensure consistency with the reference panel (strand orientation, reference allele, position. Pre-phasing was performed using Shapeit2 v2r837 [19,20] using the Shapeit2 duohmm option11[21] and cohort family structure in order to improve imputation quality [22]. Variants with low imputation quality (INFO<0.4) as well as monogenic variants were removed from the imputed set resulting in 24,111,857 variants for downstream analysis.

Ancestry assignment
Ancestry outliers were removed from the dataset. These were defined as individuals who were more than six standard deviations away from the mean in a principal component analysis of GS merged with 1092 participants from the 1000 Genomes Project[6].

Comparing two Hazard Ratios

To determine whether the hazard ratios (HRs) differ by sex, we compared the HRs directly as well as more formally in an interaction. When comparing the HRs directly we employed the following method:

Firstly the difference between the log(HRs) for men and women is calculated:

$$\text{Delta} = \log(HR_{men}) - \log(HR_{women})$$

The standard error of the difference is calculated as:

$$\text{Standard error of difference} = (SE_{men})^2 + (SE_{women})^2$$

Finally, a Z-score is created to derive a p-value:

$$\text{Z-score} = \text{Delta} / \text{Standard error of difference}$$

PGS Selection

The optimal PRS methodology is likely to vary across traits and cohorts. For example, breast cancer seems to have a slightly less polygenic architecture as PGS by the pruning and thresholding method perform better than a Bayesian method such as LDPred[23]. Whereas previous work has shown genome-wide scores are generally better than smaller scores[24]. Rather than wade into the score differences across traits, we selected a method, MegaPRS, that has outperformed other methods and should perform broadly well (as shown in Zhang et al[25]). MegaPRS

incorporates several different Bayesian methods simultaneously and was selected as a default method to generate all scores.

We aimed for the most harmonized analysis possible across all traits, by selecting the same method for all and limiting analysis to the SNPs in the intersection of HapMap phase 3 SNPs and the 1000 Genomes with a minor allele frequency greater than 1% in at least one super population (M=1,330,820). Therefore the scores across cohorts are including the same or very largely overlapping set of SNPs, making the effect sizes of scores comparable across cohorts from the same disease. From Genomics England, FinnGen, HUNT, and MGBB the percentage overlap of SNPs within each trait's PGS ranged from 83.4-100%.

MegaPRS by default generates from the full provided GWAS summary statistics two different sets called pseudo GWAS "training" and "testing" sets. The primary use of pseudo summary statistics is to construct and train prediction models, in order to decide parameters of the effect size's prior distribution[25]. Therefore no individual level data was used to train or test different versions of PGSs, however MegaPRS does internally compose several versions of PGSs together with some statistics about their predictive ability.

Future work would benefit from comparing existing scores, particularly ones being moved into clinical care, within this framework. As a comparison, the HR for the top 1% PRS versus the median (40-60%) in UK Biobank European individuals was 3.82 (3.46-4.21 95% CI) with the MegaPRS score in this study. In another publication using UK Biobank, the same comparison had an HR of 3.52 (2.93-4.24) using the Mavaddat, AJHG, 2019 score[26].

Consortia

We acknowledge the contribution of the Genomics England Research Consortium. The members of this consortium are: John C. Ambrose[1], Prabhu Arumugam[1], Roel Bevers[1], Marta Bleda[1], Freya Boardman-Pretty[1,2], Christopher R. Boustred[1], Helen Brittain[1], Matt J. Brown[1], Mark J. Caulfield[1,2], Georgia C. Chan[1], Adam Giess[1], Angela Hamblin[1], Shirley Henderson[1,2], Tim J. P. Hubbard[1], Rob Jackson[1], Louise J. Jones[1,2], Dalia Kasperaviciute[1,2], Melis Kayikci[1], Athanasios Kousathanas[1], Lea Lahnstein[1], Sarah E. A. Leigh[1], Ivonne U. S. Leong[1], Javier F. Lopez[1], Fiona Maleady-Crowe[1], Meriel McEntagart[1], Federico Minneci[1], Jonathan Mitchell[1], Loukas Moutsianas[1,2], Michael Mueller[1,2], Nirupa Murugaesu[1], Anna C. Need[1,2], Peter O'Donovan[1], Chris A.Odhams[1], Christine Patch[1,2], Daniel Perez-Gil[1], Mariana Buongermino Pereira[1], John Pullinger[1], Tahrima Rahim[1], Augusto Rendon[1], Tim Rogers[1], Kevin Savage[1], Kushmita Sawant[1], Richard H.Scott[1], Afshan Siddiq[1], Alexander Sieghart[1], Samuel C. Smith[1], Alona Sosinsky[1,2], Alexander Stuckey[1], Mélanie Tanguy[1], Ana Lisa Taylor Tavares[1], Ellen R. A. Thomas[1,2], Simon R.Thompson[1], Arianna Tucci[1,2], Matthew J. Welland[1], Eleanor Williams[1], Katarzyna Witkowska[1,2], Suzanne M. Wood[1,2], Magdalena Zarowiecki[1].
1. Genomics England, London, UK
2. William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, London, UK

We acknowledge the contribution of the FinnGen Research project. The members of the team, alphabetized by first name then last name, are:
Aarno Palotie[1,2], Adam Ziemann[3], Adriana Huertas-Vazquez[4], Aino Salminen[5], Airi Jussila[6], Aki Havulinna[7], Alessandro Porello[8], Amanda Elliott[9,10], Amy Hart[8], Anastasia Kytölä[2], Anders Mälarstig[11], Andrea Ganna[2], Andrey Loboda[4], Anna Vlahiotis[11], Anne Lehtonen[3], Anne Remes[12], Annika Auranen[13], Antti Aarnisalo[5], Antti Hakanen[14], Antti Mäkitie[15], Antti Palomäki[16], Anu Jalanko[2], Anu Loukola[17], Aoxing Liu[2], Apinya Lertratanakul[3], Argyro Bizaki-Vallaskangas[6], Arto Lehisto[2], Arto Mannermaa[18], Åsa Hedman[11], Audrey Chu[19], Auli Toivola[20], Awaisa Ghazal[2], Bridget Riley-Gillis[3], Chia-Yen Chen[21], Chris O´Donnell[22], Clément Chatelain[23], Coralie Viollet[24], Daniel Gordin[5], David Choy[25], David Pulford[26], David Rice[5], Dawn Waterworth[8], Debby Ngo[27], Denise Öller[2], Dermot Reilly[28], Diptee Kulkarni[19], Dirk Paul[29], Edmond Teng[25], Eero Punkka[17], Eeva Kangasniemi[30], Eeva Sliz[31], Eija Laakkonen[32], Ekaterina Khramtsova[8], Elina Järvensivu[20], Elina Kilpeläinen[2], Elisa Rahikkala[12], Elisabeth Widen[2], Elmo Saarentaus[2], Erich Strauss[25], Erkki Isometsä[5], Esa Pitkänen[2], Essi Kaiharju[20], Eveliina Salminen[5], Fabiana Farias[4], Fanli Xu[19], Fedik Rahimov[3], Felix Vaura[33], Fredrik Åberg[34], George Okafo[35], Hanati Tuoken[35], Hanna Ollila[2], Hannele Laivuori[2], Hannele Mattsson[20], Hannu Kankaanranta[36], Hannu Uusitalo[6], Hao Chen[25], Harri Siirtola[37], Heidi Silven[31], Heikki Joensuu[38], Heiko Runz[21], Helen Cooper[2], Heli Lehtonen[11], Henna Palin[30], Henrike Heyne[2], Hilkka Soininen[39], Howard Jacob[3], Hubert Chen[25], Huei-Yi Shen[2], Iida Vähätalo[37], Iiris Hovatta[40], Ilkka Kalliala[5], Ioanna Tachmazidou[29], Jaakko Kaprio[2], Jaakko Parkkinen[11], Jaakko Tyrmi[41], Jaana Suvisaari[33], Jae-Hoon Sul[4], Janet Kumar[42], Jani Tikkanen[43], Jari Laukkanen[44,45], Jarmo Ritari[46], Jason Miller[4], Javier Garcia-Tabuenca[37], Jeffrey Waring[3], Jenni Aittokallio[16], Jennifer Schutzman[25], Jiwoo Lee[47], Joanna Betts[19], Joel Rämö[2], Johanna Huhtakangas[12], Johanna Mäkelä[48], Johanna Mattson[5], Johanna Schleutker[49,16], Johannes Kettunen[43,12], John Eicher[19], Jonas Zierer[22], Jonathan Chung[22], Joni A Turunen[50], Jorge Esparza Gordillo[19], Joseph Maranville[51], Juha Karjalainen[2], Juha Mehtonen[2], Juha Rinne[16], Juha Sinisalo[5], Jukka Koskela[5], Jukka Partanen[52], Jukka Peltola[6], Jussi Hernesniemi[6], Juulia Partanen[53], Jyrki Tammerluoto[2], Kai Kaarniranta[54], Kaisa Tasanen[12], Kaj Metsärinne[16], Karen He[8], Kari Eklund[5], Karol Estrada[55], Katariina Hannula-Jouppi[5], Katherine Klinger[23], Kati Donner[2], Kati Hyvärinen[46], Kati Kristiansson[20], Katja Kivinen[2], Katri Kaukinen[6], Katri Pylkäs[31], Katriina Aalto-Setälä[56], Kimmo Palin[57], Kirsi Auro[58], Kirsi Kalpala[11], Kirsi Sipilä[59], Klaus Elenius[16], Kristiina Aittomäki[60], Kristin Tsuo[47], L. Elisa Lahtela[2], Laura Addis[19], Laura Huilaja[12], Laura Kotaniemi-Talonen[6], Laura Pirilä[16], Laure

Morin-Papunen[12], Lauri Aaltonen[5], Leena Koulu[16], Liisa Suominen[39], Lila Kallio[14], Linda McCarthy[19], Lotta Männikkö[20], Ma´en Obeidat[22], Maarit Niinimäki[12], Majd Mouded[27], Malla-Maria Linna[17], Manuel Rivas[61], Marc Jung[35], Marco Hautalahti[62], Margaret G. Ehm[42], Margarete Fabre[24], Margit Pelkonen[39], Mari Kaunisto[2], Mari Niemi[22], Maria Siponen[39], Marianna Niemi[37], Marja Vääräsmäki[12], Marja-Riitta Taskinen[5], Mark Daly[2,1], Mark McCarthy[25], Markku Laukkanen[20], Markku Voutilainen[16], Markus Perola[20], Marla Hochfeld[51], Martti Färkkilä[5], Mary Pat Reeve[2], Masahiro Kanai[63], Matthias Gossel[23], Meijian Guan[8], Melissa Miller[11], Mengzhen Liu[3], Mervi Aavikko[2], Mika Helminen[37], Mika Kähönen[30,6], Mike Mendelson[64], Mikko Arvas[52], Mikko Hiltunen[65], Mikko Kiviniemi[39], Minna Brunfeldt[20], Minna Karjalainen[31], Minna Raivio[5], Minna Ruddock[66], Mitja Kurki[47], Mutaamba Maasha[63], Nan Bing[11], Natalia Pujol[67], Natalie Bowers[25], Neha Raghavan[4], Nicole Renaud[22], Niko Välimäki[57], Nina Mars[2], Nina Pitkänen[14], Nizar Smaoui[3], Oili Kaipiainen-Seppänen[39], Olli Carpén[68,5], Oluwaseun Alexander[2], Oskari Heikinheimo[5], Outi Tuovila[69], Outi Uimari[12], Päivi Auvinen[39], Päivi Laiho[20], Päivi Mäntylä[39], Paula Kauppi[5], Peeter Karihtala[38], Pekka Nieminen[5], Pentti Tienari[5], Petri Lehto[62], Petri Virolainen[14], Pia Isomäki[6], Pietro Della Briotta Paralo[2], Pirkko Pussinen[5], Priit Palta[2], Qingqin S Li[70], Raimo Pakkanen[69], Raisa Serpi[43], Rajashree Mishra[19], Rasko Leinonen[71], Reetta Hinttala[43], Reetta Kälviäinen[39], Regis Wong[20], Relja Popovic[3], Rigbe Weldatsadik[2], Riikka Arffman[31], Riitta Lahesmaa[16], Rion Pendergrass[25], Risto Kajanne[2], Robert Graham[55], Robert Plenge[51], Rodos Rodosthenous[2], Roosa Kallionpää[16], Sally John[21], Sami Heikkinen[65], Sami Koskelainen[20], Sampsa Pikkarainen[5], Samuel Lessard[23], Samuli Ripatti[2], Sanna Siltanen[30], Sanna Toppila-Salmi[72], Sanni Lahdenperä[21], Sanni Ruotsalainen[2], Sarah Smith[62], Satu Strausz[2], Sauli Vuoti[73], Shabbeer Hassan[2], Shameek Biswas[51], Shanmukha Sampath Padmanabhuni[2], Shuang Luo[2], Simonne Longerich[4], Sini Lähteenmäki[20], Sirkku Peltonen[16], Slavé Petrovski[29], Stefan McDonough[11], Stephanie Loomis[21], Susan Eaton[21], Susanna Lemmelä[2], Susanna Savukoski[31], Taneli Raivio[17], Tarja Laitinen[30,6], Taru Tukiainen[2], Teea Salmi[6], Teemu Niiranen[33], Teemu Paajanen[20], Teijo Kuopio[44], Terhi Kilpi[20], Terhi Ollila[5], Terhi Piltonen[12], Tero Hiekkalinna[20], Terttu Harju[12], Tiina Luukkaala[37], Tiina Wahlfors[20], Tiinamaija Tuomi[5], Tim Lu[25], Timo Blomster[12], Timo Hiltunen[5], Timo P. Sipilä[2], Tom Southerington[62], Tomi P. Mäkelä[74], Triin Laisk[67], Tuomo Kiiskinen[2], Tuomo Mantere[43], Tuomo Meretoja[75], Tuula Palotie[76], Tuula Salo[5], Tuuli Sistonen[20], Tytti Willberg[16], Ulla Palotie[5], Ulvi Gursoy[16], Valtteri Julkunen[39], Varpu Jokimaa[16], Veikko Salomaa[33], Veli-Matti Kosma[18], Venla Kurra[6], Vincent Llorens[2], Vuokko Anttonen[12], Wei Zhou[63], Xinli Hu[11], Ying Wu[11], Zhihao Ding[35], Zhili Zheng[63]

[1]Broad Institute of MIT and Harvard; Massachusetts General Hospital, [2]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland, [3]Abbvie, Chicago, IL, United States, [4]Merck, Kenilworth, NJ, United States, [5]Hospital District of Helsinki and Uusimaa, Helsinki, Finland, [6]Pirkanmaa Hospital District, Tampere, Finland, [7]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland; Finnish Institute for Health and Welfare (THL), Helsinki, Finland, [8]Janssen Research & Development, LLC, Spring House, PA, United States, [9]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland; Broad Institute, Cambridge, MA, USA and Massachusetts General Hospital, Boston, MA, USA, [10]Broad Institute, Cambridge, MA, USA and Massachusetts General Hospital, Boston, MA, USA, [11]Pfizer, New York, NY, United States, [12]Northern Ostrobothnia Hospital District, Oulu, Finland, [13]Pirkanmaa Hospital District , Tampere, Finland, [14]Auria Biobank / University of Turku / Hospital District of Southwest Finland, Turku, Finland, [15]Department of Otorhinolaryngology - Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland, [16]Hospital District of Southwest Finland, Turku, Finland, [17]Helsinki Biobank / Helsinki University and Hospital District of Helsinki and Uusimaa, Helsinki, [18]Biobank of Eastern Finland / University of Eastern Finland / Northern Savo Hospital District, Kuopio, Finland, [19]GlaxoSmithKline, Brentford, United Kingdom, [20]THL Biobank / Finnish Institute for Health and Welfare (THL), Helsinki, Finland, [21]Biogen, Cambridge, MA, United States, [22]Novartis Institutes for BioMedical Research, Cambridge, MA, United States, [23]Translational Sciences, Sanofi R&D, Framingham, MA, USA,

[24]AstraZeneca, Cambridge, United Kingdom, [25]Genentech, San Francisco, CA, United States, [26]GlaxoSmithKline, Stevenage, United Kingdom, [27]Novartis, Basel, Switzerland, [28]Janssen Research & Development, LLC, Boston, MA, United States, [29]Astra Zeneca, Cambridge, United Kingdom, [30]Finnish Clinical Biobank Tampere / University of Tampere / Pirkanmaa Hospital District, Tampere, Finland, [31]University of Oulu, Oulu, Finland, [32]University of Jyväskylä, Jyväskylä, Finland, [33]Finnish Institute for Health and Welfare (THL), Helsinki, Finland, [34]Transplantation and Liver Surgery Clinic, Helsinki University Hospital, Helsinki University, Helsinki, Finland, [35]Boehringer Ingelheim, Ingelheim am Rhein, Germany, [36]University of Gothenburg, Gothenburg, Sweden/ Seinäjoki Central Hospital, Seinäjoki, Finland/ Tampere University, Tampere, Finland, [37]University of Tampere, Tampere, Finland, [38]Department of Oncology, Helsinki University Hospital Comprehensive Cancer Center and University of Helsinki, Helsinki, Finland, [39]Northern Savo Hospital District, Kuopio, Finland, [40]University of Helsinki, Finland, [41]University of Oulu, Oulu, Finland / University of Tampere, Tampere, Finland, [42]GlaxoSmithKline, Collegeville, PA, United States, [43]Northern Finland Biobank Borealis / University of Oulu / Northern Ostrobothnia Hospital District, Oulu, Finland, [44]Central Finland Biobank / University of Jyväskylä / Central Finland Health Care District, Jyväskylä, Finland, [45]Central Finland Health Care District, Jyväskylä, Finland, [46]Finnish Red Cross Blood Service, Helsinki, Finland, [47]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland; Broad Institute, Cambridge, MA, United States, [48]FINBB - Finnish biobank cooperative, [49]Auria Biobank / Univ. of Turku / Hospital District of Southwest Finland, Turku, Finland, [50]Helsinki University Hospital and University of Helsinki, Helsinki, Finland; Eye Genetics Group, Folkhälsan Research Center, Helsinki, Finland, [51]Bristol Myers Squibb, New York, NY, United States, [52]Finnish Red Cross Blood Service / Finnish Hematology Registry and Clinical Biobank, Helsinki, Finland, [53]Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland, [54]Northern Savo Hospital District, Kuopio, Finland; Department of Molecular Genetics, University of Lodz, Lodz, Poland, [55]Maze Therapeutics, San Francisco, CA, United States, [56]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, [57]University of Helsinki, Helsinki, Finland, [58]GlaxoSmithKline, Espoo, Finland, [59]Research Unit of Oral Health Sciences Faculty of Medicine, University of Oulu, Oulu, Finland; Medical Research Center, Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland, [60]Department of Medical Genetics, Helsinki University Central Hospital, Helsinki, Finland, [61]University of Stanford, Stanford, CA, United States, [62]Finnish Biobank Cooperative - FINBB, [63]Broad Institute, Cambridge, MA, United States, [64]Novartis, Boston, MA, United States, [65]University of Eastern Finland, Kuopio, Finland, [66]Arctic biobank / University of Oulu, [67]Estonian biobank, Tartu, Estonia, [68]Helsinki Biobank / Helsinki University, [69]Business Finland, Helsinki, Finland, [70]Janssen Research & Development, LLC, Titusville, NJ 08560, United States, [71]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland; European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK, [72]University of Eastern Finland and Kuopio University Hospital, Department of Otorhinolaryngology, Kuopio, Finland and Department of Allergy, Helsinki University Hospital and University of Helsinki, Finland, [73]Janssen-Cilag Oy, Espoo, Finland, [74]HiLIFE, University of Helsinki, Finland, Finland, [75]Department of Breast Surgery, Helsinki University Hospital Comprehensive Cancer Center and University of Helsinki, Helsinki, Finland, [76]University of Helsinki and Hospital District of Helsinki and Uusimaa, Helsinki, Finland

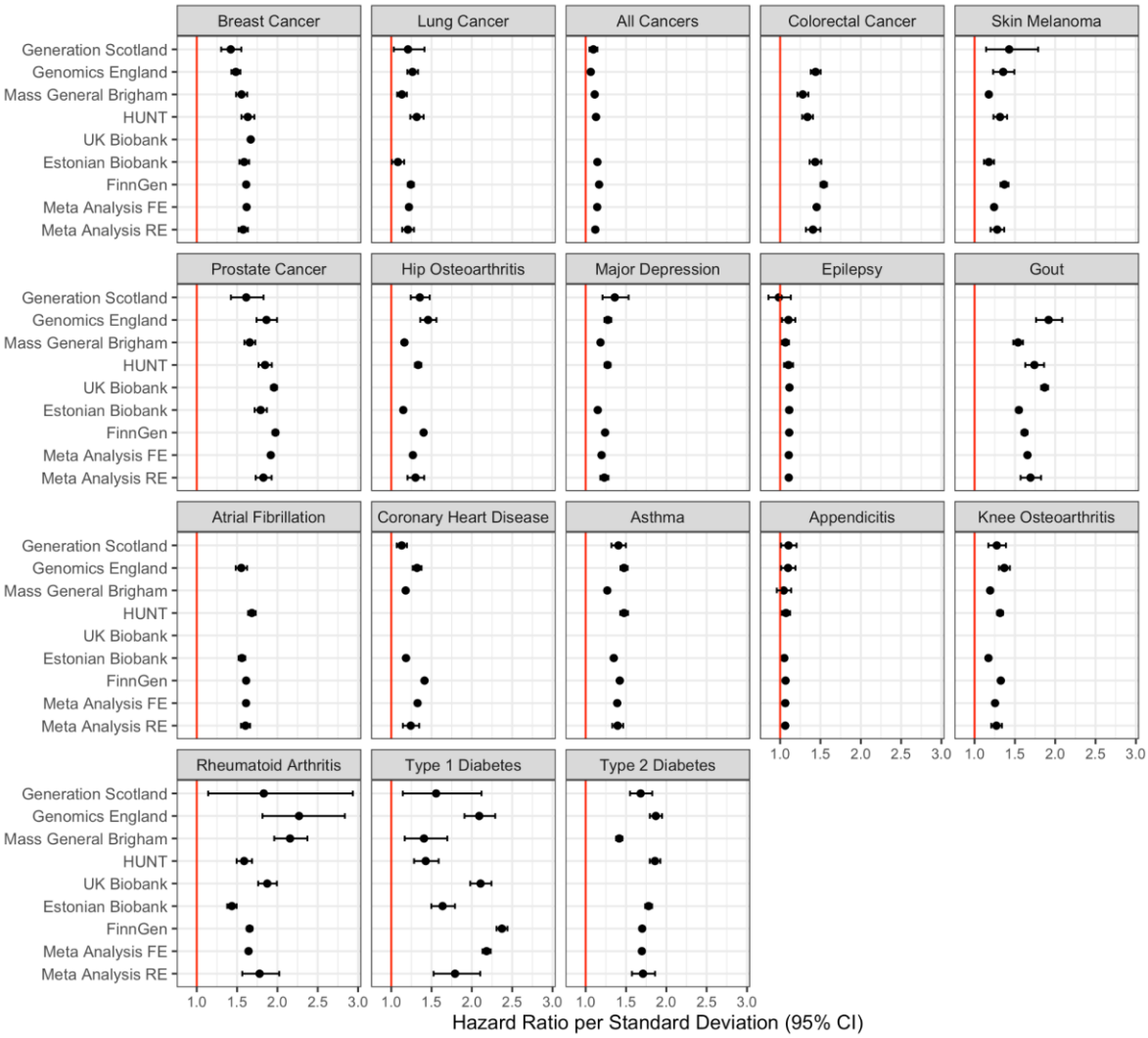
We acknowledge the contribution of the Estonian Biobank research team. The members of the team are:
Andres Metspalu[1,2], Lili Milani[1], Tõnu Esko[1], Reedik Mägi,[1] Mari Nelis[3], Georgi Hudjashov[1].
1. Institute of Genomics, University of Tatu, Estonia
2. The Institute of Molecular and Cell Biology, University of Tartu, Estonia
3. Core Facility of Genomics, University of Tartu, Estonia

The current members of HUNT All-In Research Team (in alphabetical order by surname):

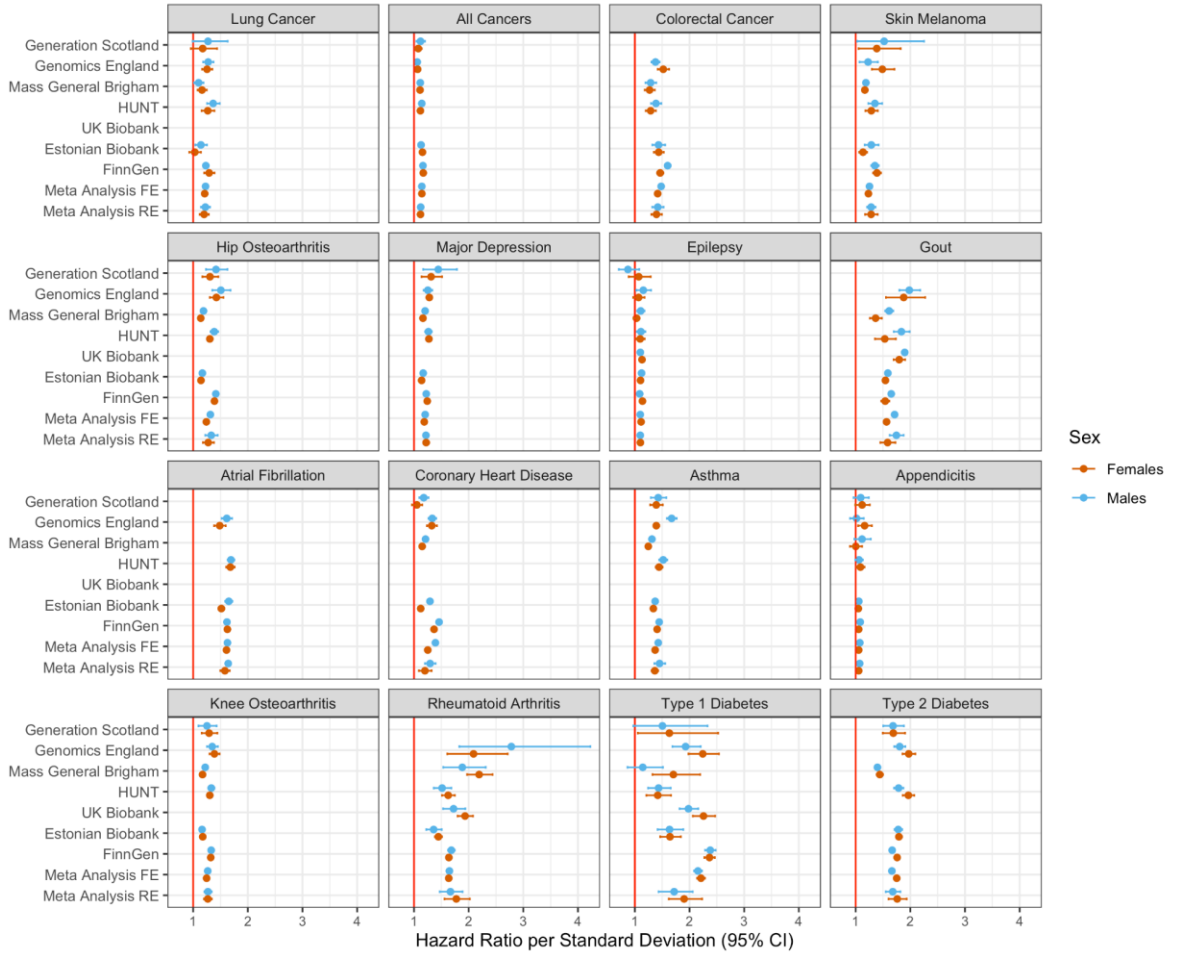
Bjørn Olav Åsvold, Ben Brumpton, Maiken Elvestad Gabrielsen, Kristian Hveem, Ida Surakka, Laurent Thomas, Wei Zhou

Supplementary Figures

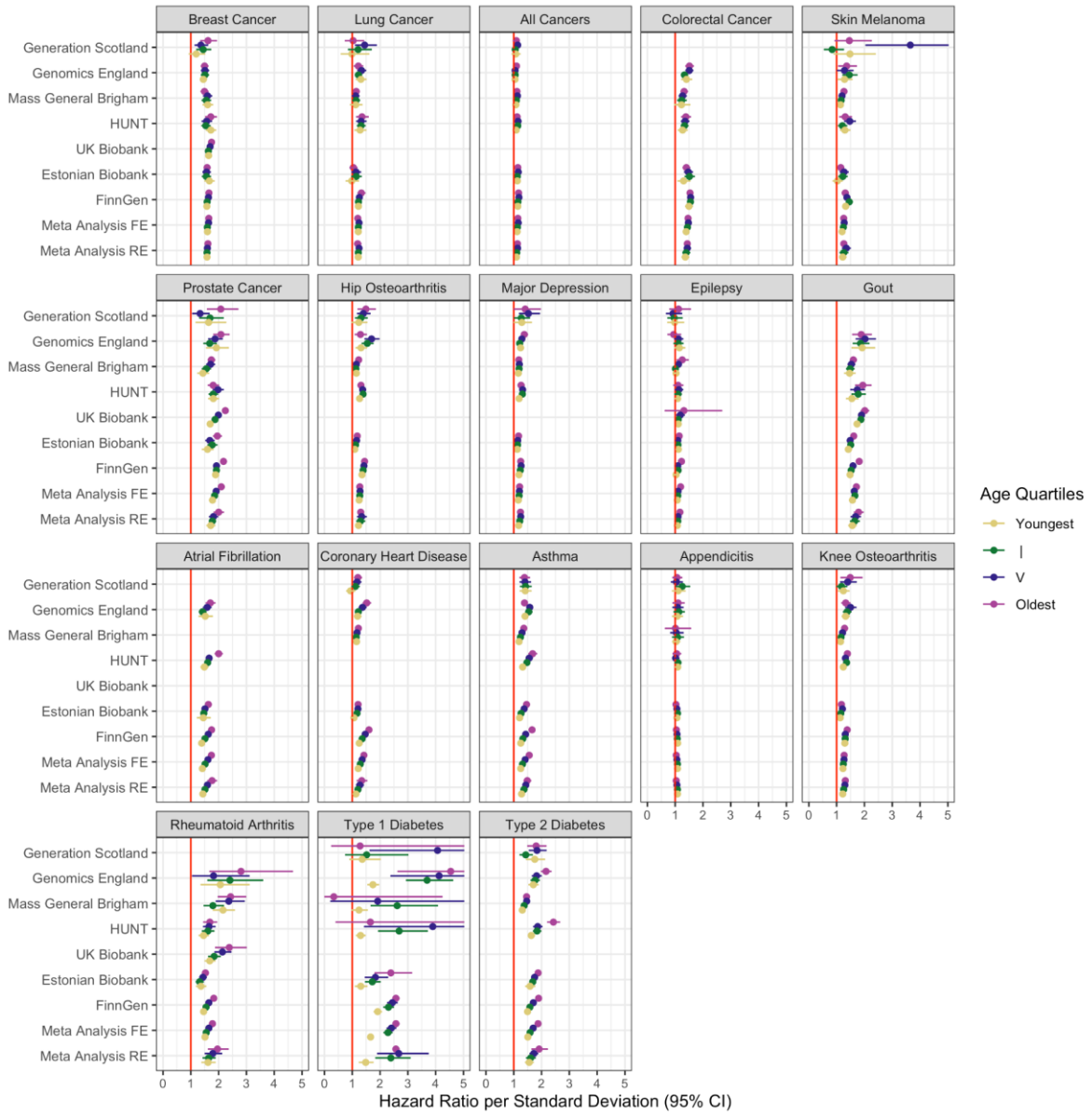
a)



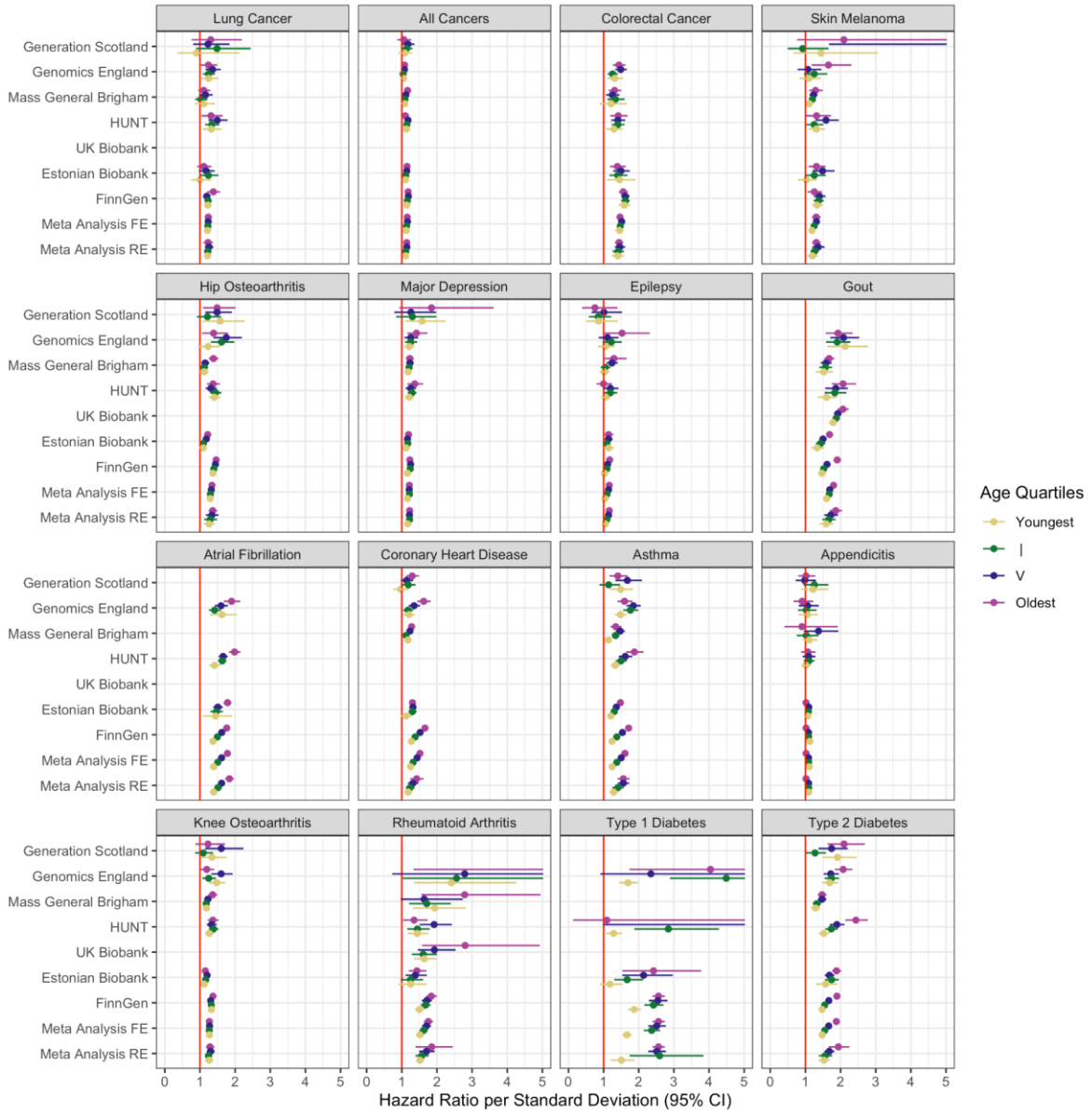
b)


Lung Cancer
All Cancers
Colorectal Cancer
Skin Melanoma
Hip Osteoarthritis
Major Depression
Epilepsy
Gout
Atrial Fibrillation
Coronary Heart Disease
Asthma
Appendicitis
Knee Osteoarthritis
Rheumatoid Arthritis
Type 1 Diabetes
Type 2 Diabetes
Generation Scotland
Genomics England
Mass General Brigham
HUNT
UK Biobank
Estonian Biobank
FinnGen
Meta Analysis FE
Meta Analysis RE
1
2
3
4
Hazard Ratio per Standard Deviation (95% CI)
Sex
Females
Males

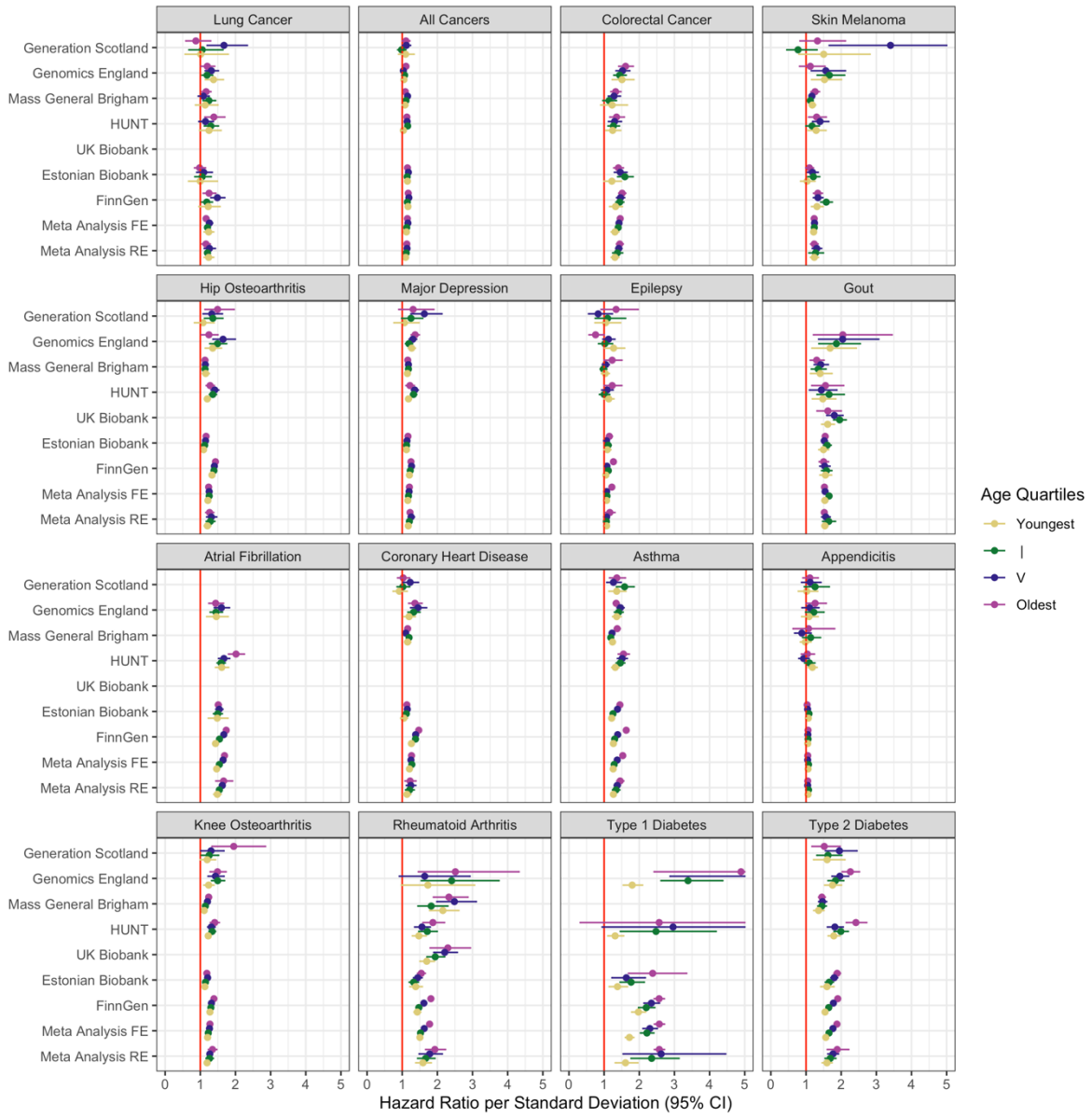
c)


Breast Cancer
Lung Cancer
All Cancers
Colorectal Cancer
Skin Melanoma
Prostate Cancer
Hip Osteoarthritis
Major Depression
Epilepsy
Gout
Atrial Fibrillation
Coronary Heart Disease
Asthma
Appendicitis
Knee Osteoarthritis
Rheumatoid Arthritis
Type 1 Diabetes
Type 2 Diabetes
Generation Scotland
Genomics England
Mass General Brigham
HUNT
UK Biobank
Estonian Biobank
FinnGen
Meta Analysis FE
Meta Analysis RE
0 1 2 3 4 5
Hazard Ratio per Standard Deviation (95% CI)
Age Quartiles
Youngest
I
V
Oldest

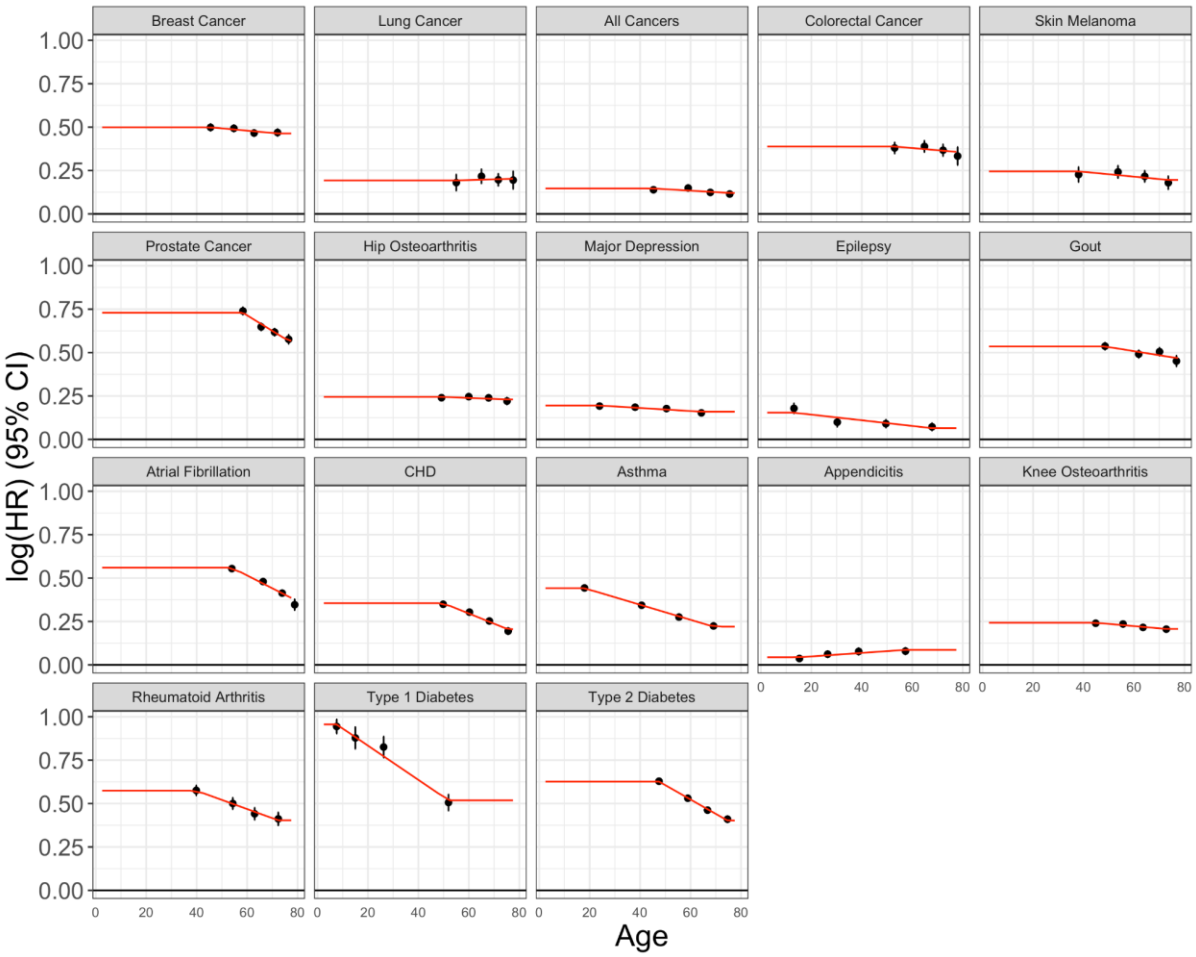
d)


Lung Cancer
All Cancers
Colorectal Cancer
Skin Melanoma
Hip Osteoarthritis
Major Depression
Epilepsy
Gout
Atrial Fibrillation
Coronary Heart Disease
Asthma
Appendicitis
Knee Osteoarthritis
Rheumatoid Arthritis
Type 1 Diabetes
Type 2 Diabetes
Generation Scotland
Genomics England
Mass General Brigham
HUNT
UK Biobank
Estonian Biobank
FinnGen
Meta Analysis FE
Meta Analysis RE
0
1
2
3
4
5
Hazard Ratio per Standard Deviation (95% CI)
Age Quartiles
Youngest
I
V
Oldest

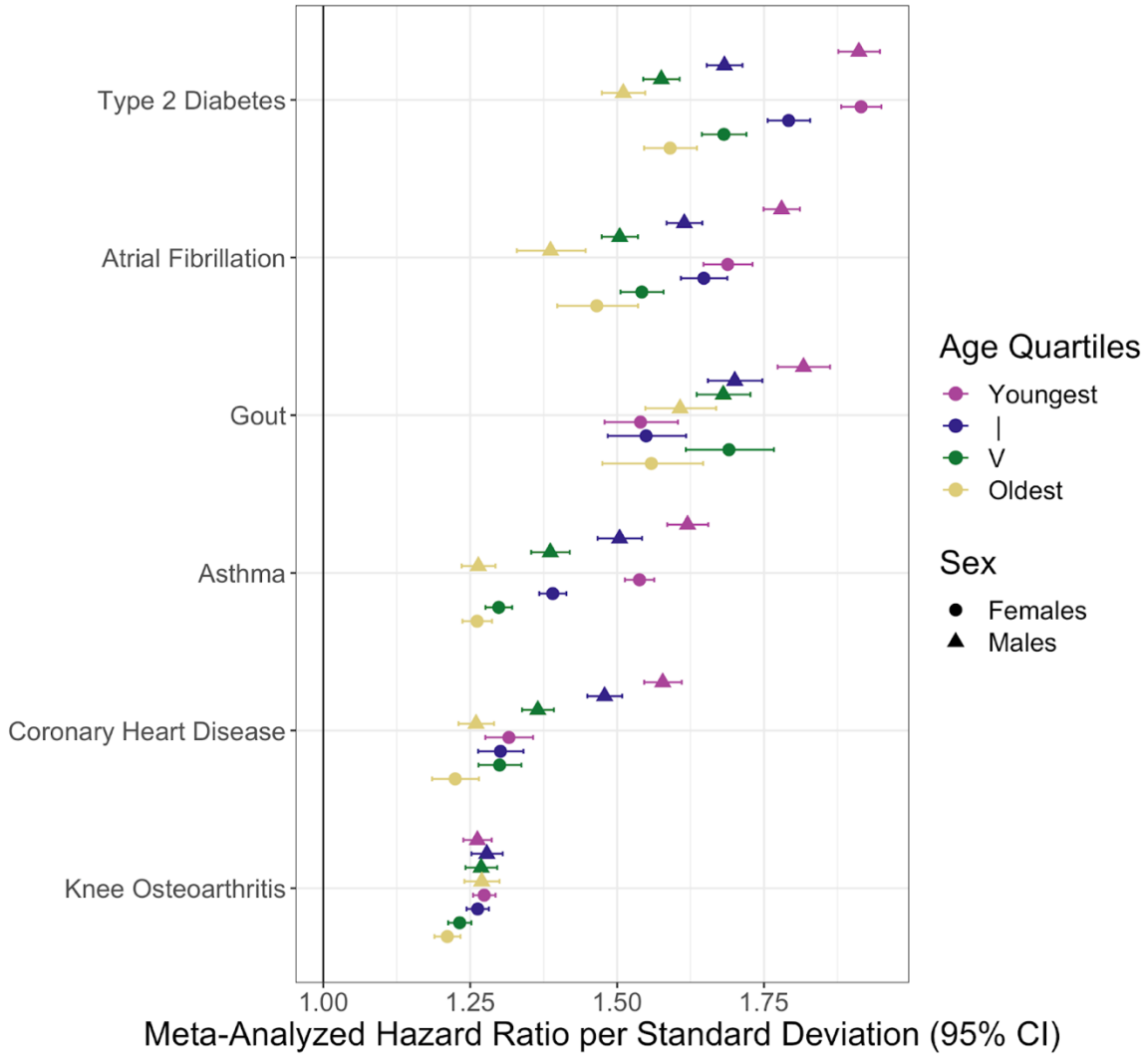
e)



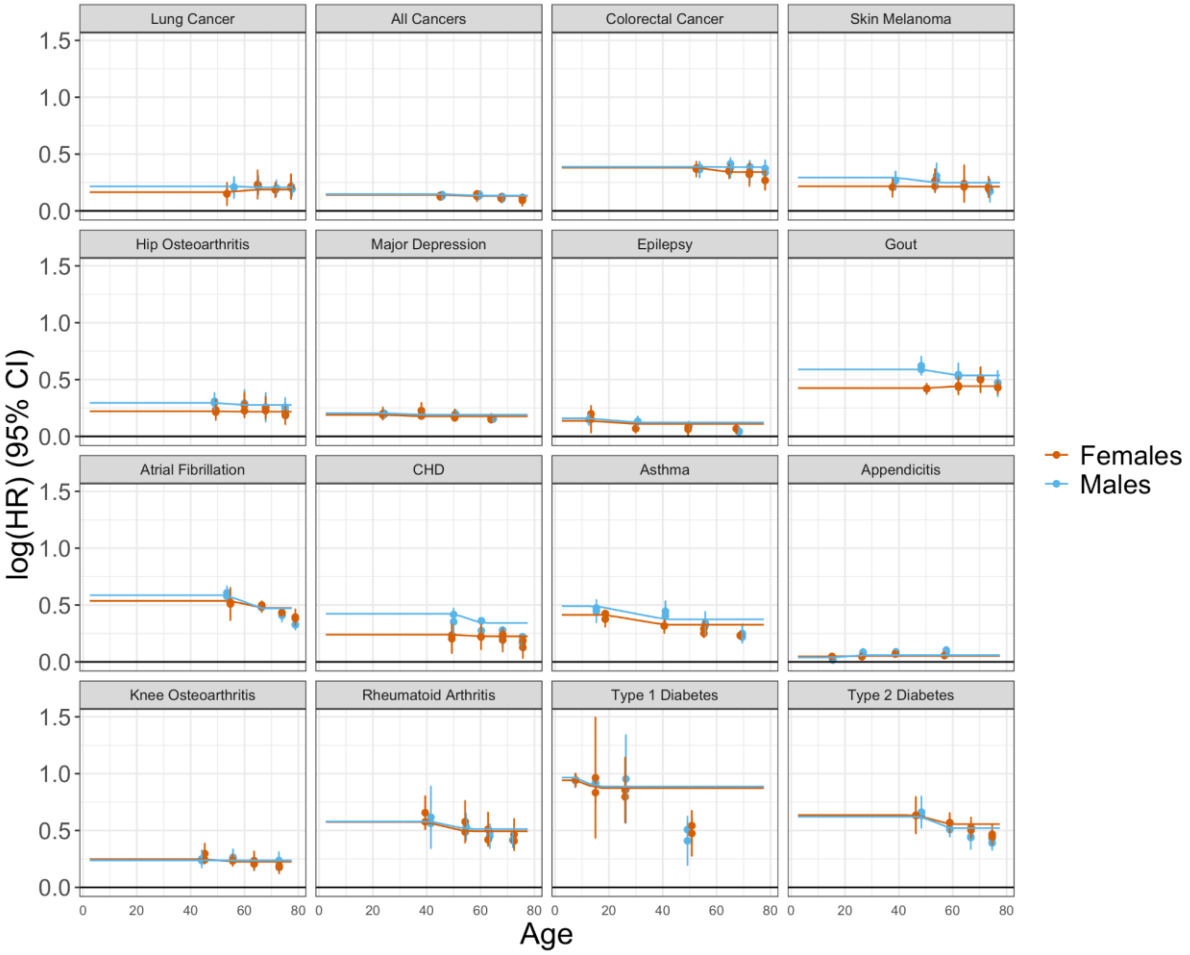
Supplementary Figure 1. Forest plots of each phenotype. a) No stratification. b) Sex stratification. c) Age stratification. d) Age and sex stratification - Males. e) Age and sex stratification - Females. Note: for figures d and e the limit has been restricted to 5 for presentability, meaning the confidence intervals has been truncated on the graphs.



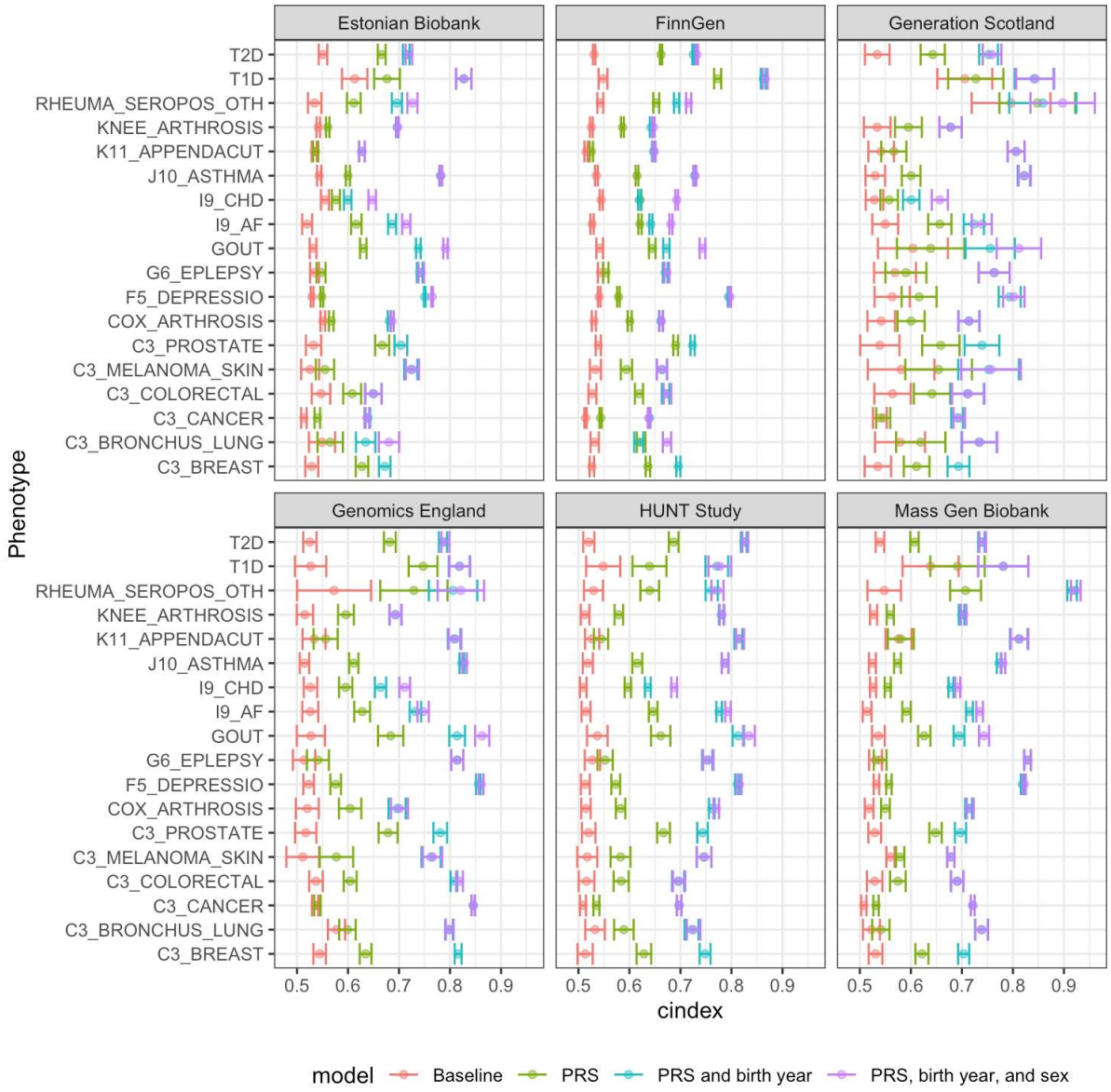
Supplementary Figure 2. Age specific effects of meta-analyzed log(Hazard Ratios)



Supplementary Figure 3. Hazard ratios per standard deviation stratified by age and sex. Note: phenotypes were only included if the model selected was age and sex stratified.

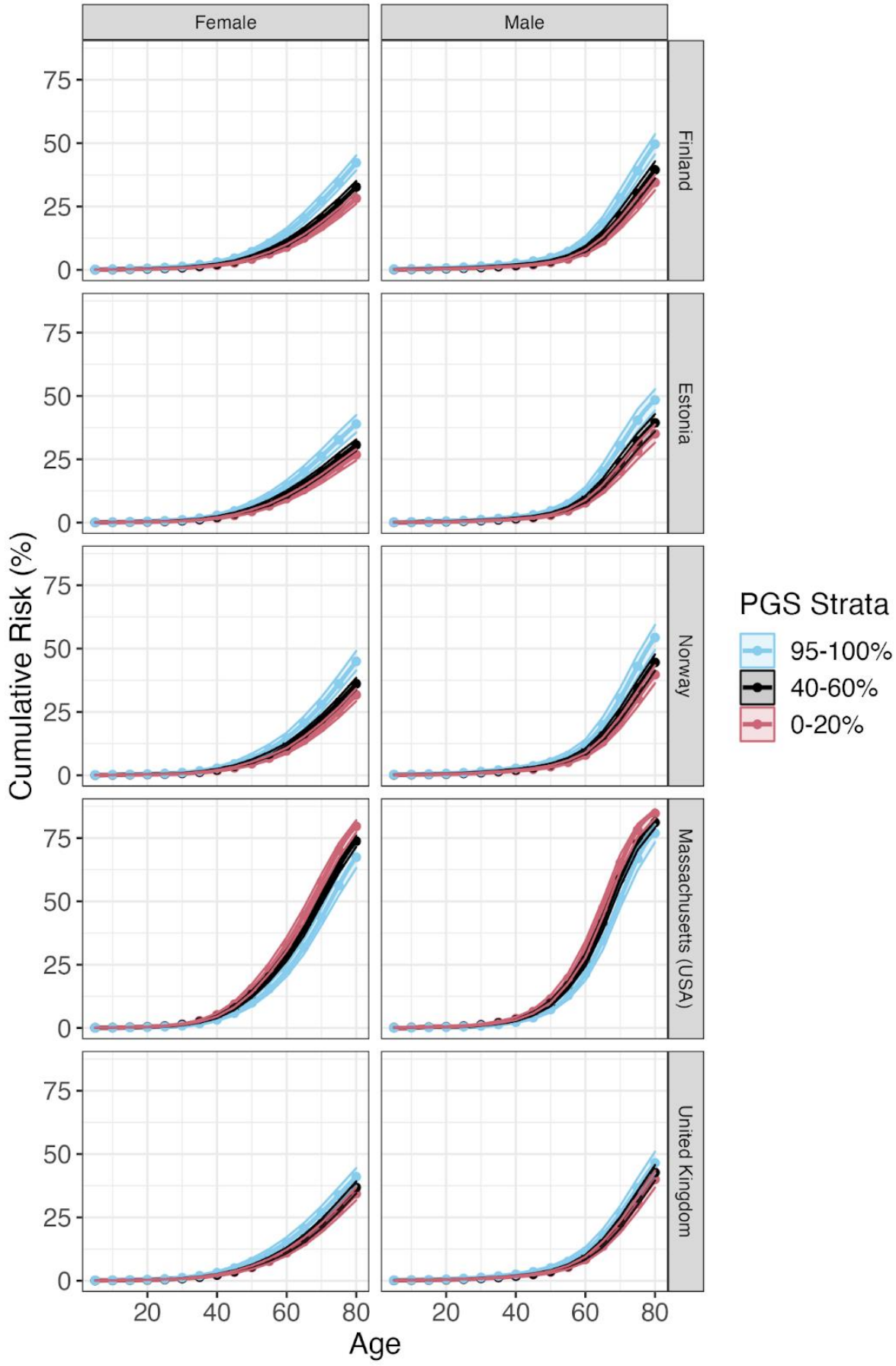


Supplementary Figure 4. Age specific effects of meta-analyzed log(Hazard Ratios) stratified by sex

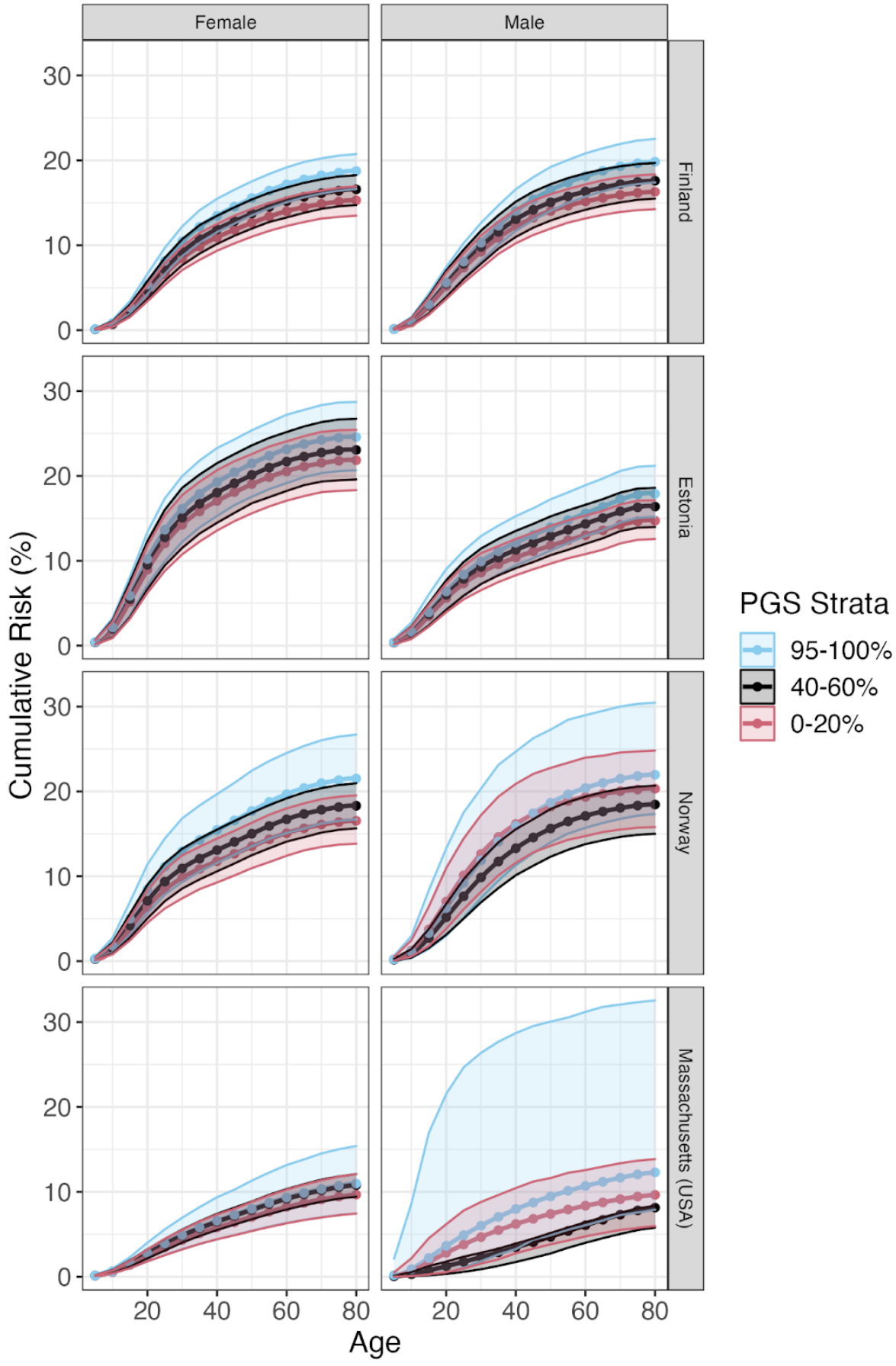


Supplementary Figure 5. The baseline model is the first 10 principal components of genetic data. The PRS as a predictor is the PRS percentile grouping, but resulst are similar when PRS is a continuous variable. Birth year was used as a proxy for age, since age was used as the time-scale of the Cox model and not as a predictor. For the model considered best for each phenotype, see Supplementary Table 13. Error bars are 95% confidence intervals for the C-statistic.

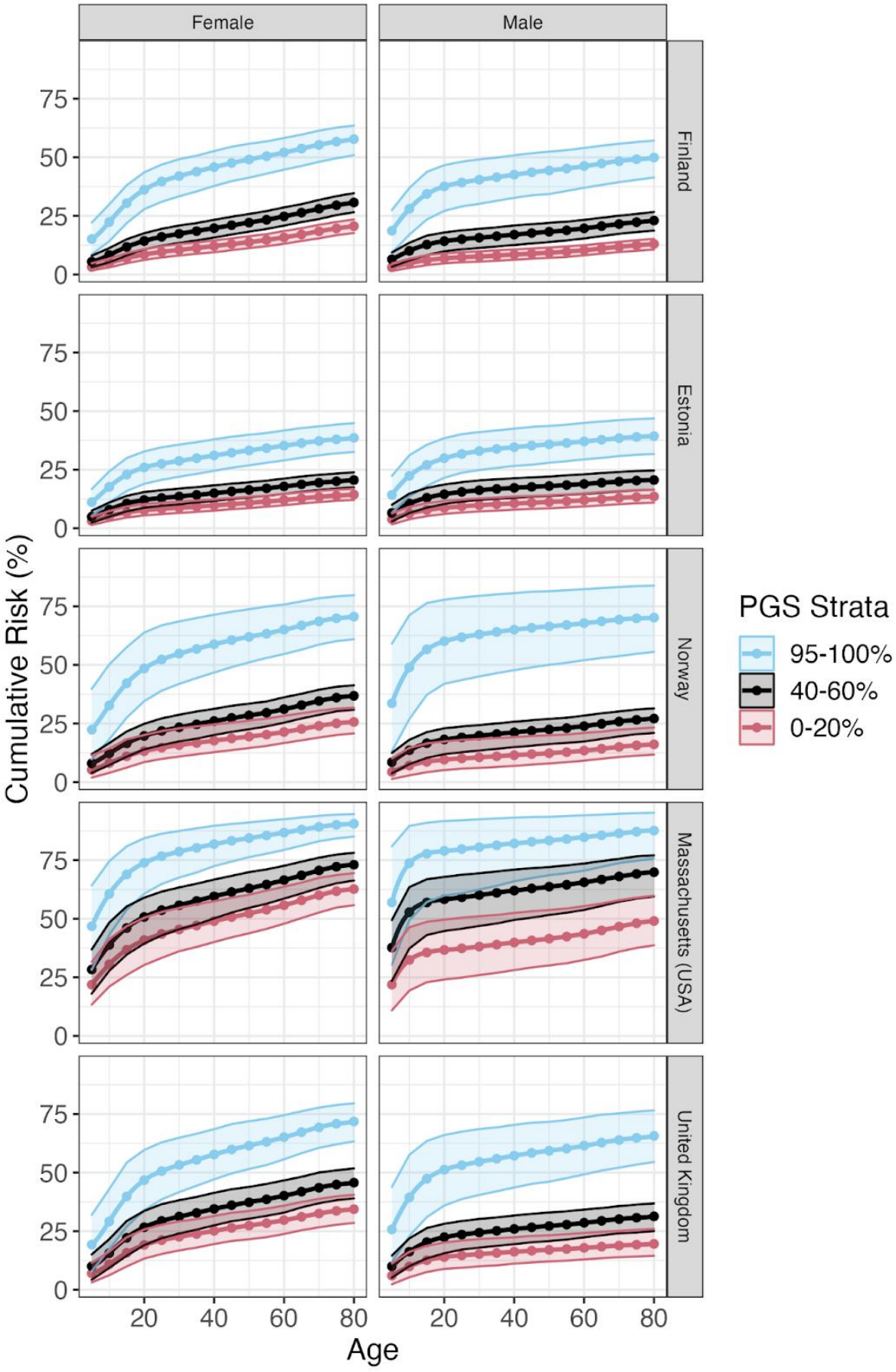
a. All Cancers



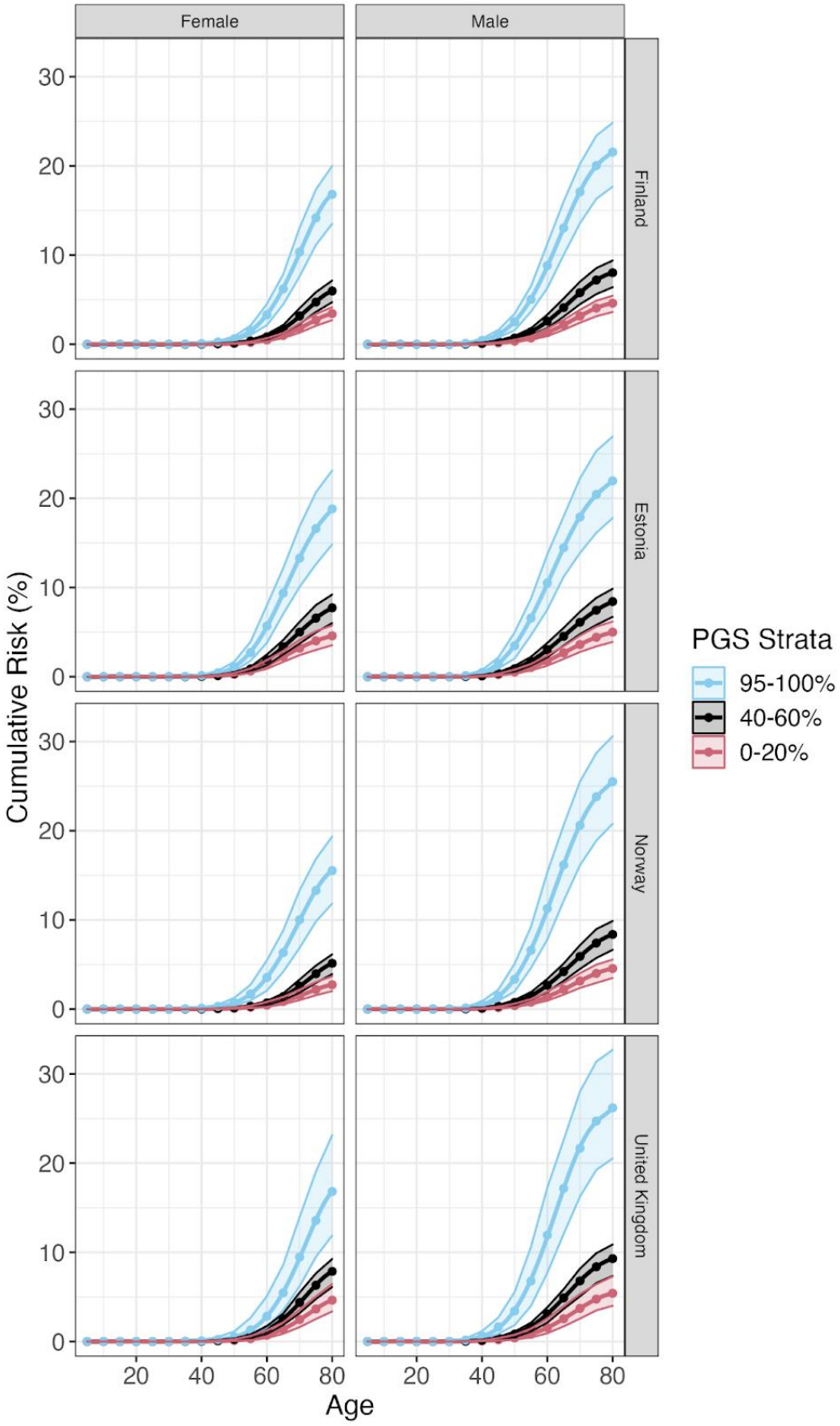
b. Appendicitis



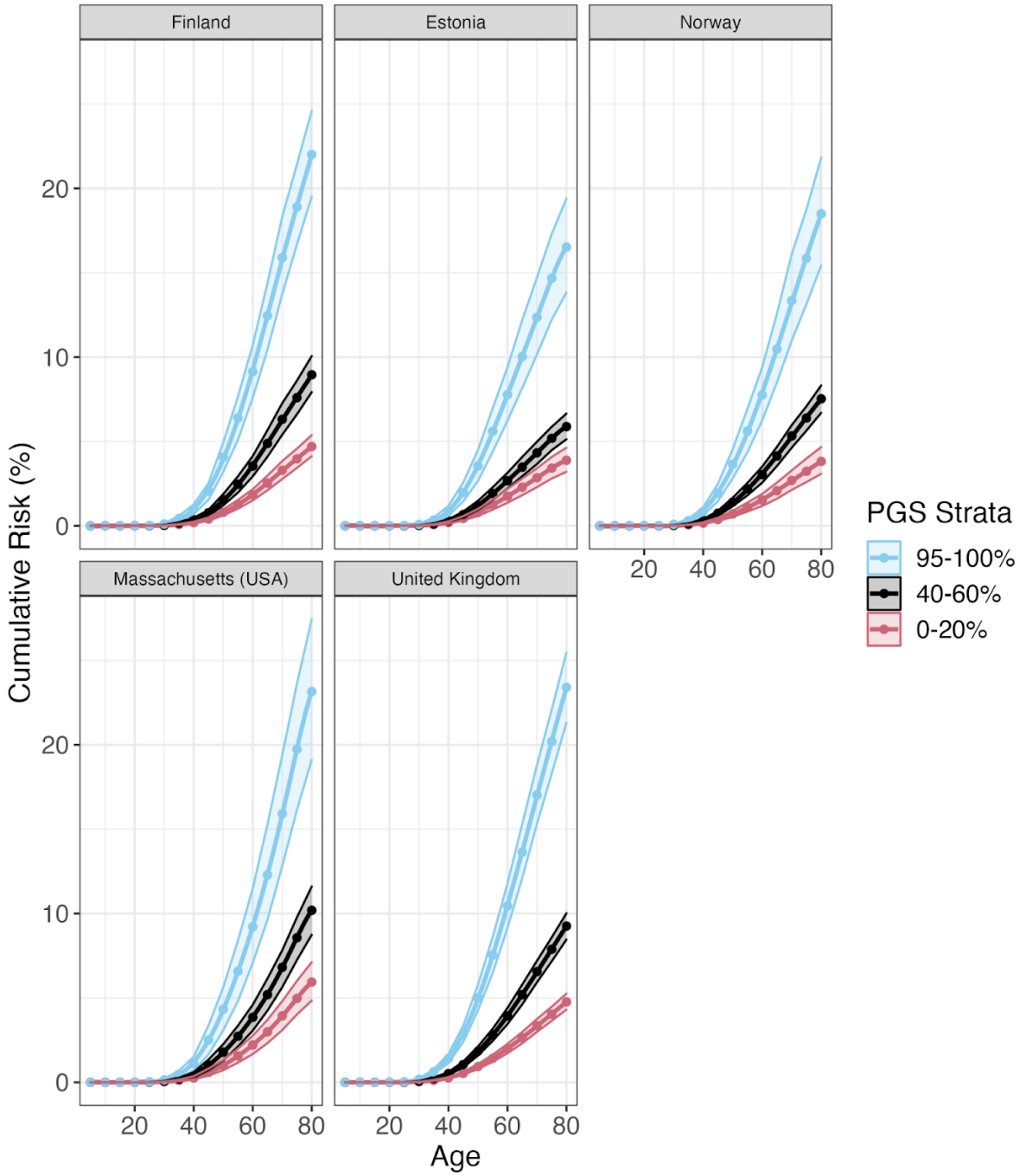
c. Asthma


Female
Male
Finland
Estonia
Norway
Massachusetts (USA)
United Kingdom
Cumulative Risk (%)
0
25
50
75
20
40
60
80
Age
PGS Strata
95-100%
40-60%
0-20%

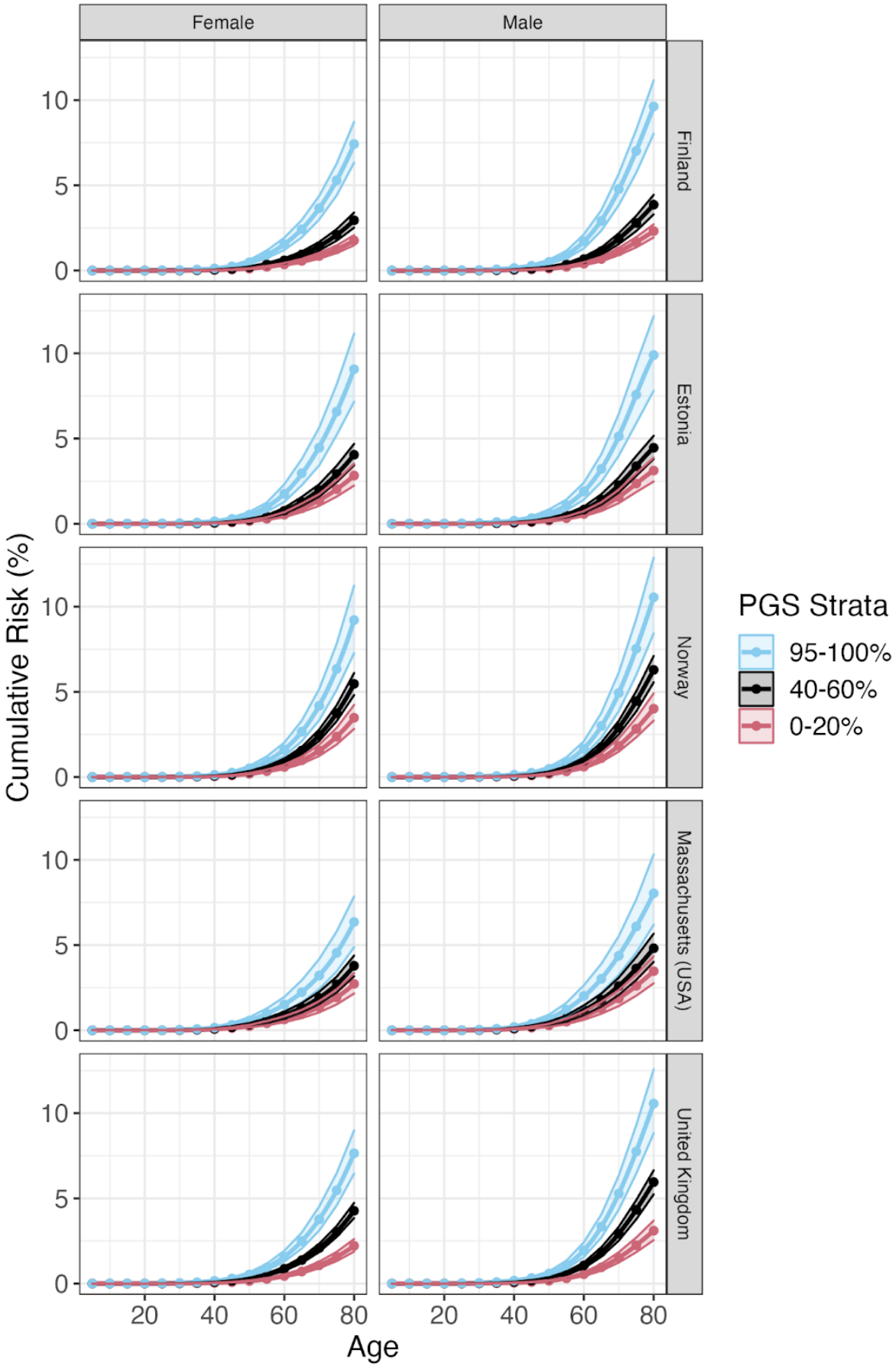
d. Atrial Fibrillation



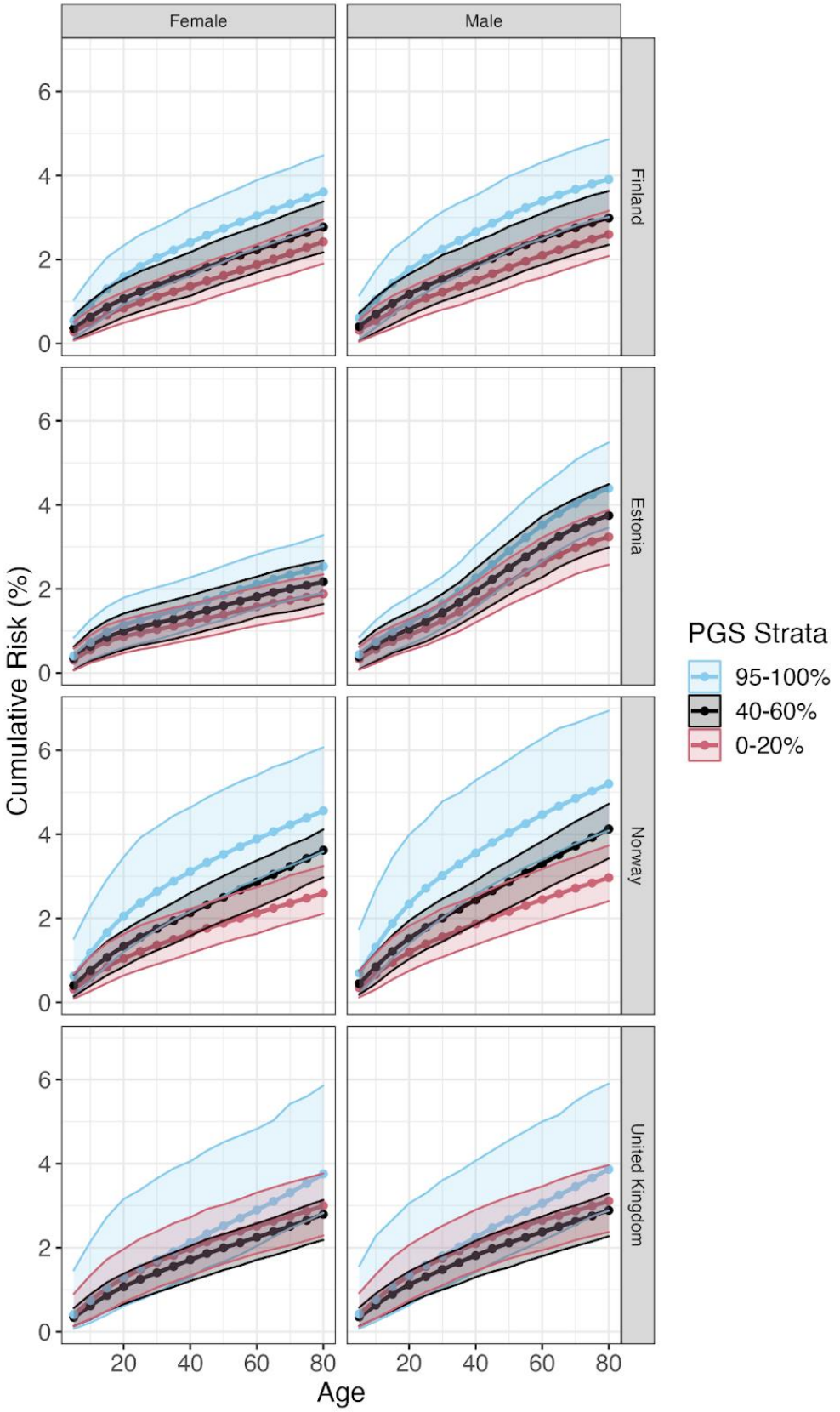
e. Breast Cancer



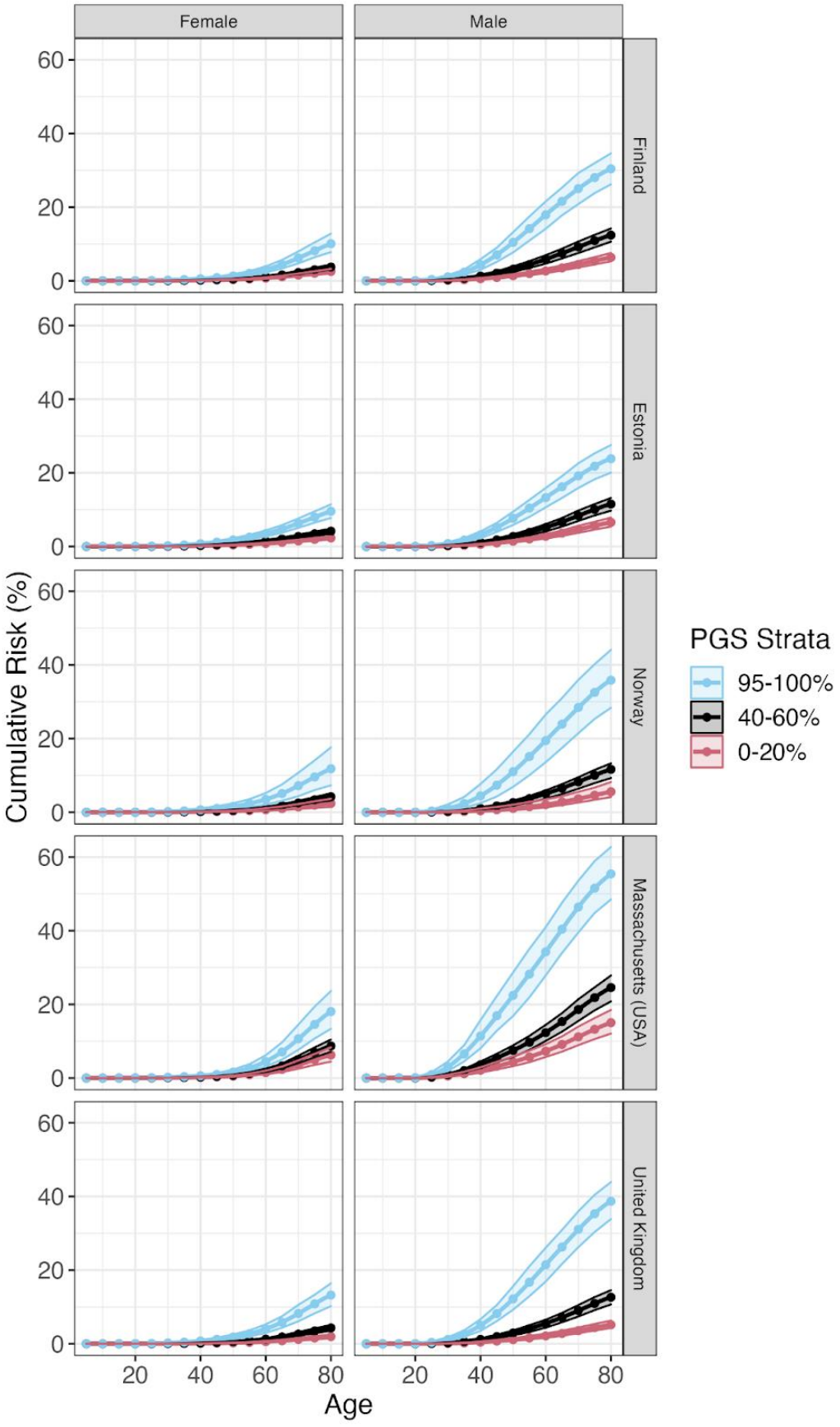
f. Colorectal Cancer



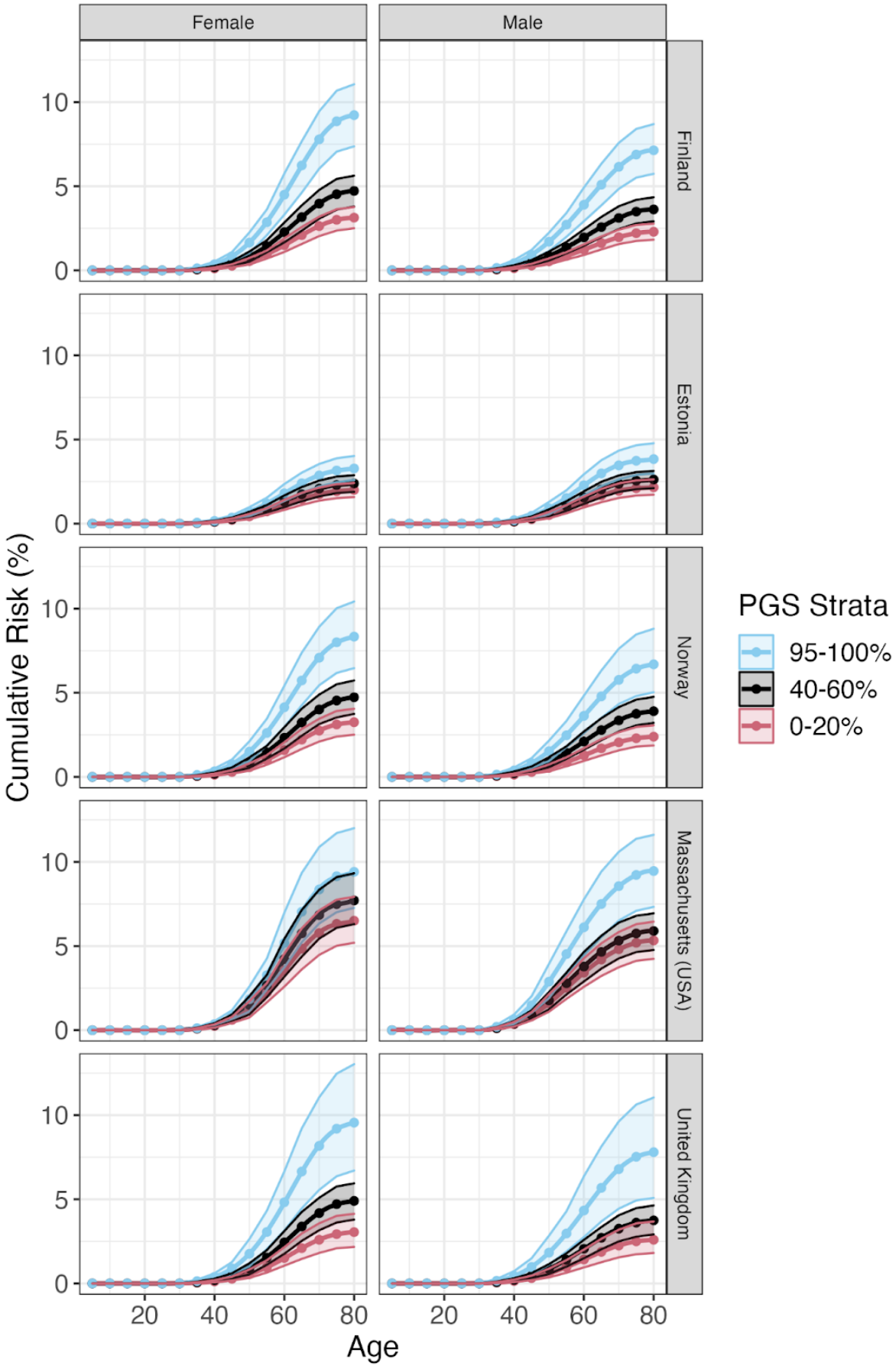
g. Epilepsy


Female
Male
Finland
Estonia
Norway
United Kingdom
Cumulative Risk (%)
0
2
4
6
20
40
60
80
Age
PGS Strata
95-100%
40-60%
0-20%

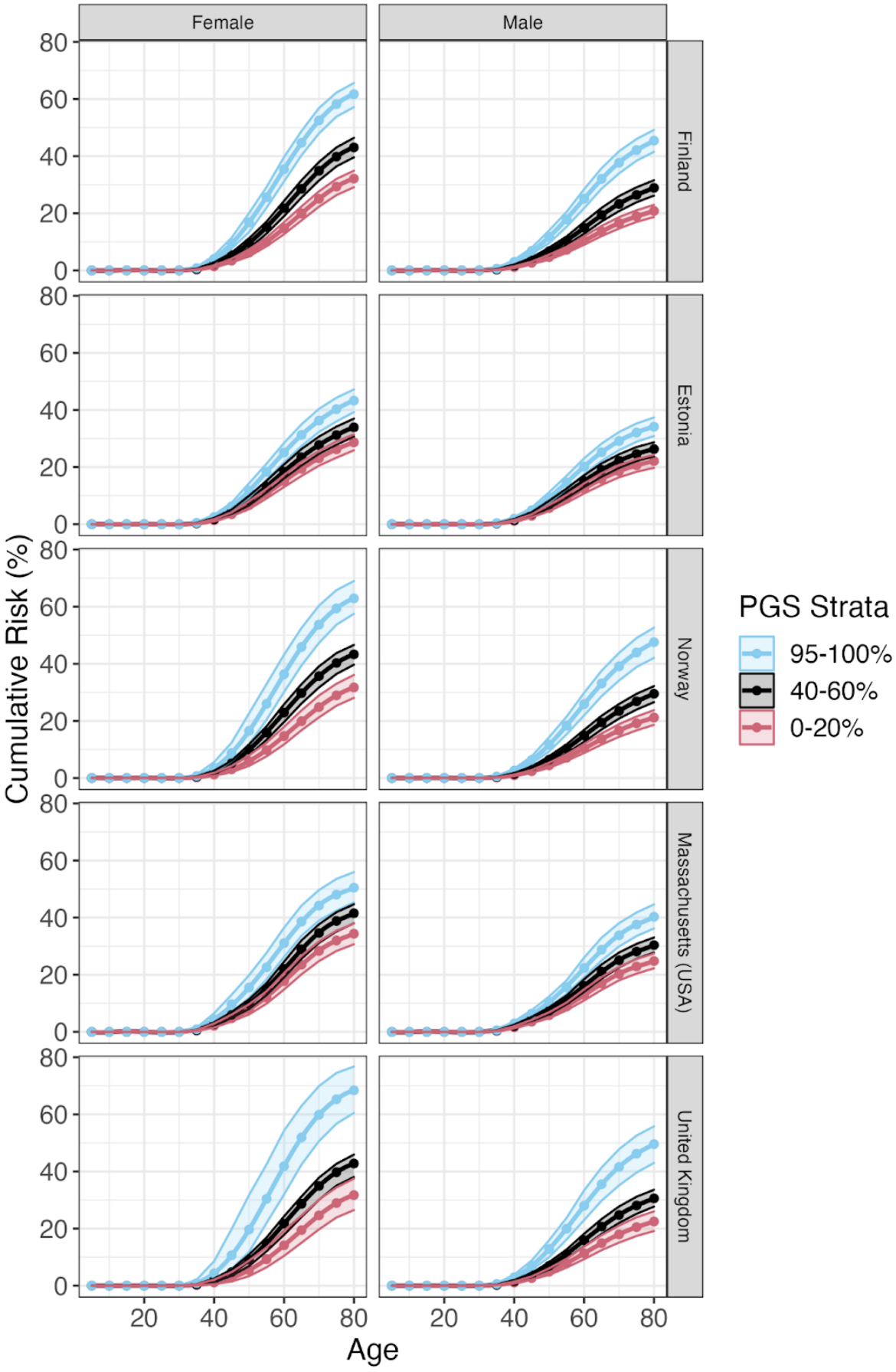
h. Gout


Female
Male
Finland
Estonia
Norway
Massachusetts (USA)
United Kingdom
Cumulative Risk (%)
0
20
40
60
20
40
60
80
Age
PGS Strata
95-100%
40-60%
0-20%

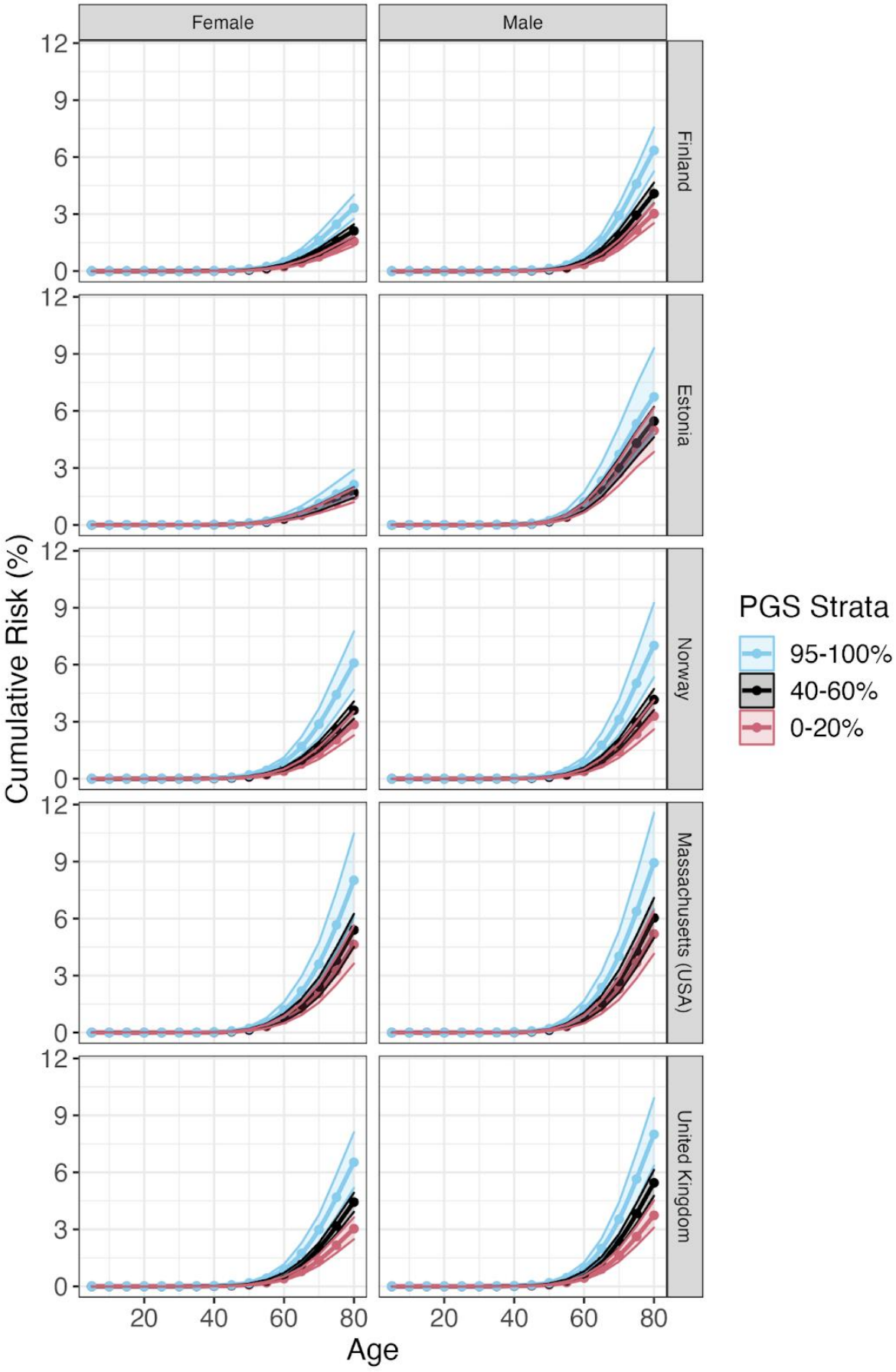
i. Hip Osteoarthritis



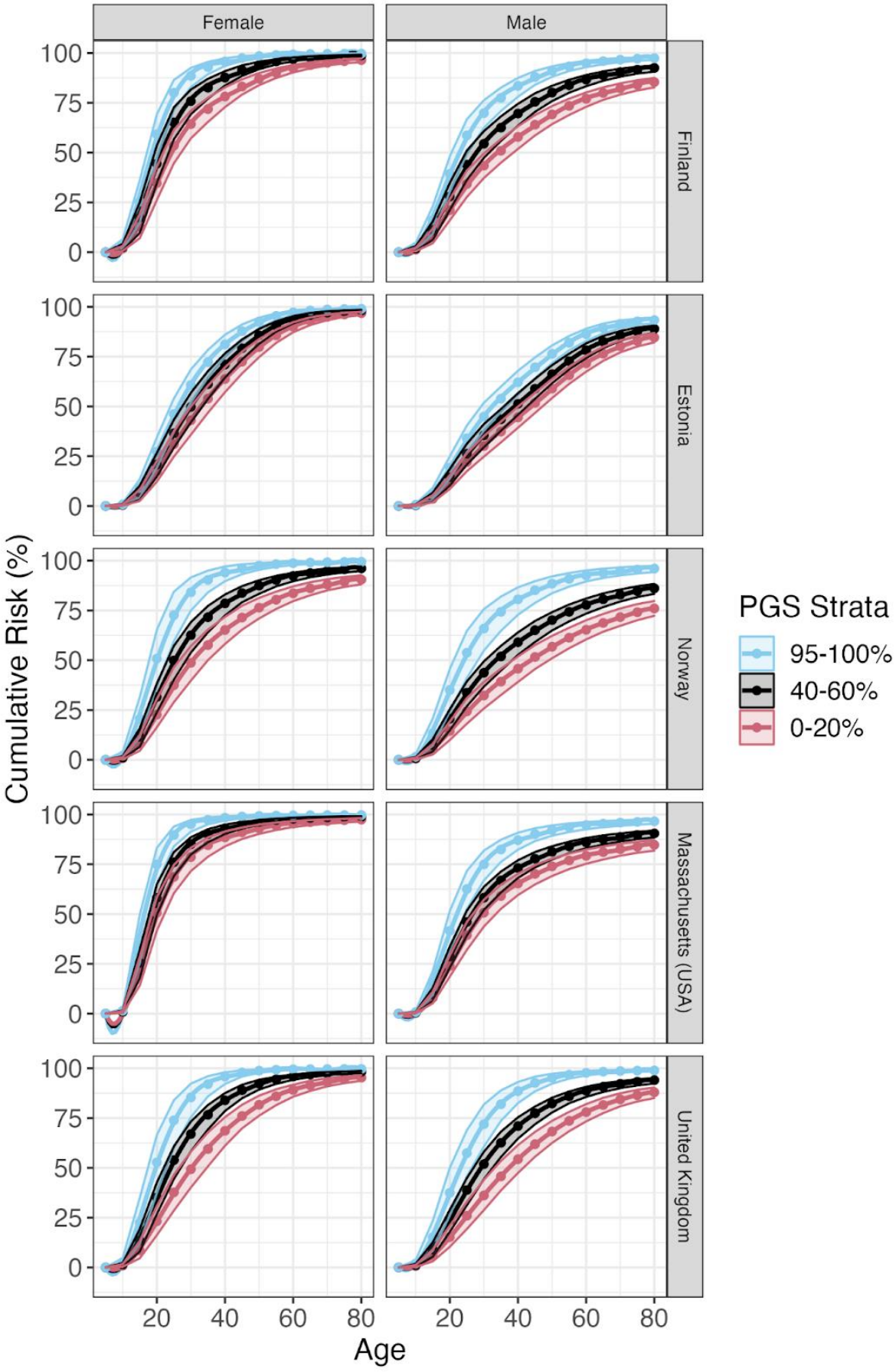
j. Knee Osteoarthritis



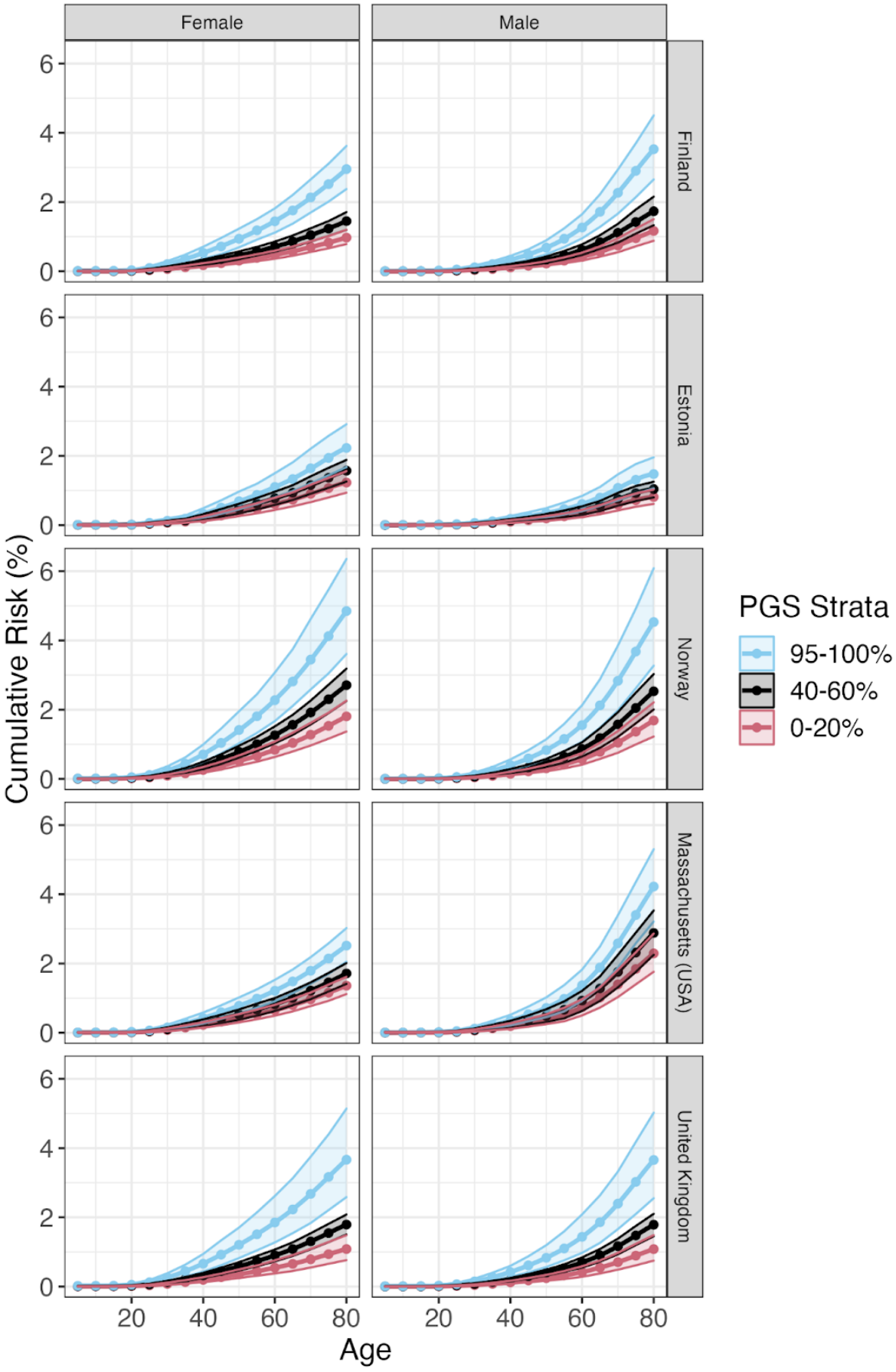
k. Lung Cancer



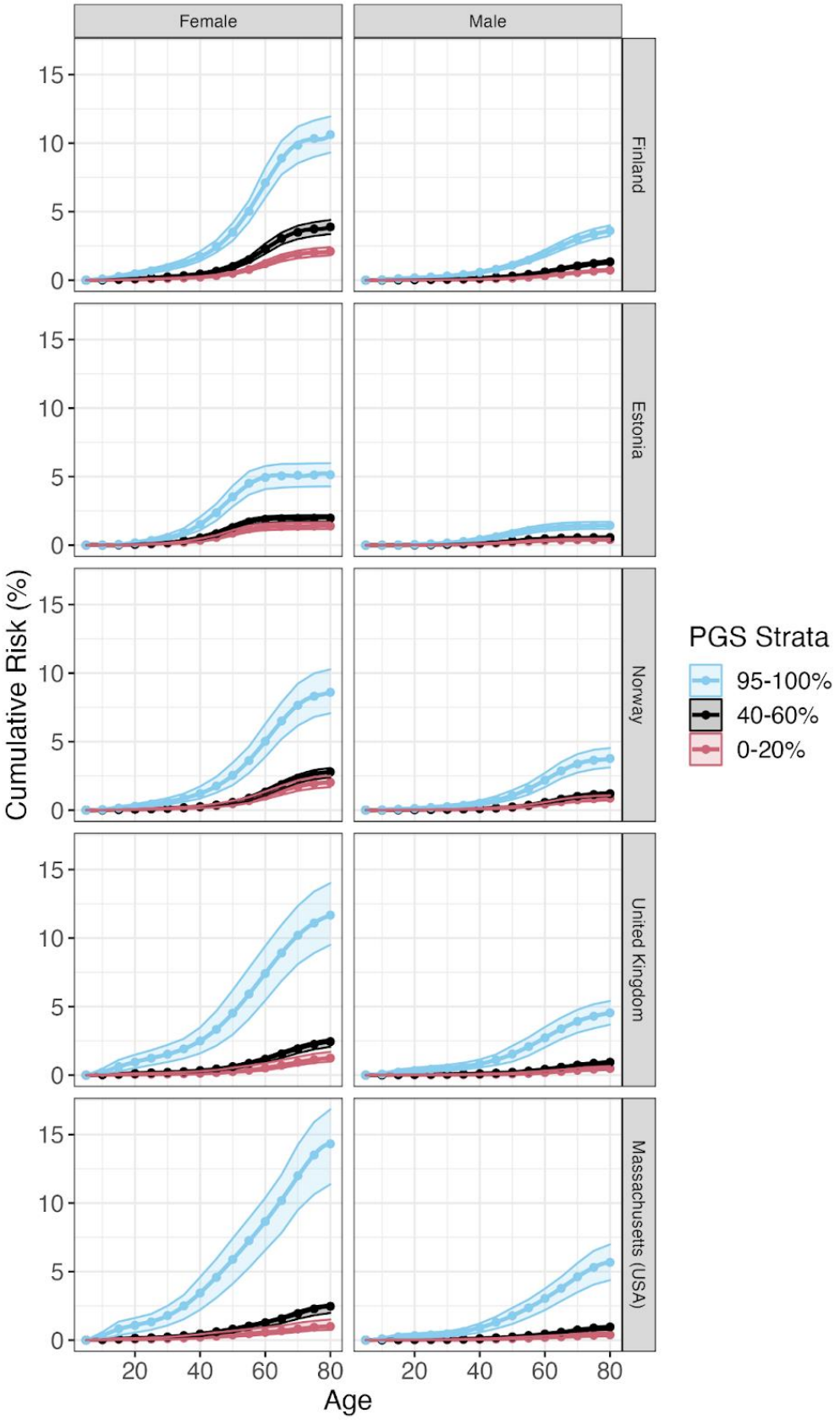
I. Depression



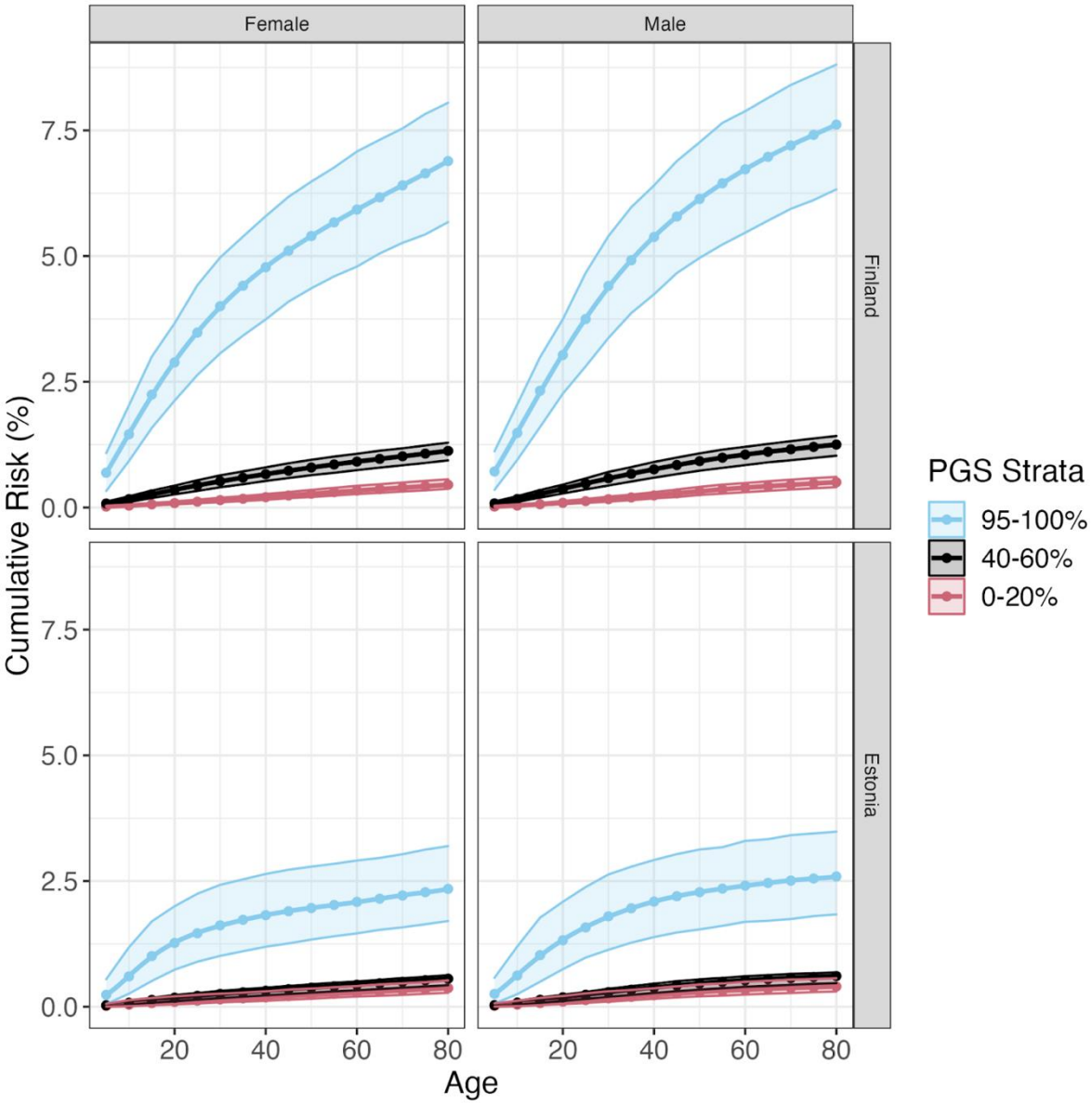
m. Skin Melanoma



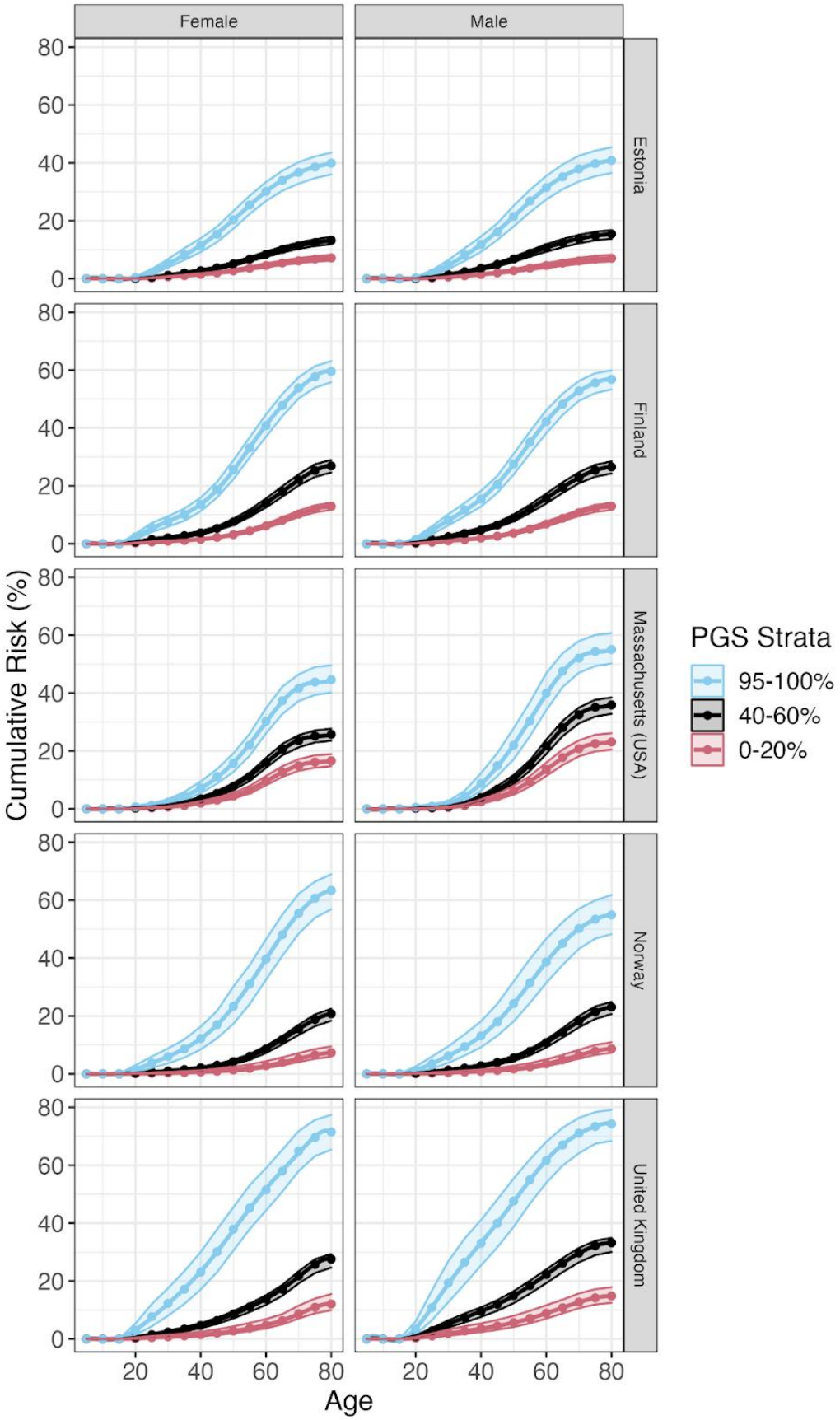
n. Rheumatoid Arthritis



o. Type 1 Diabetes

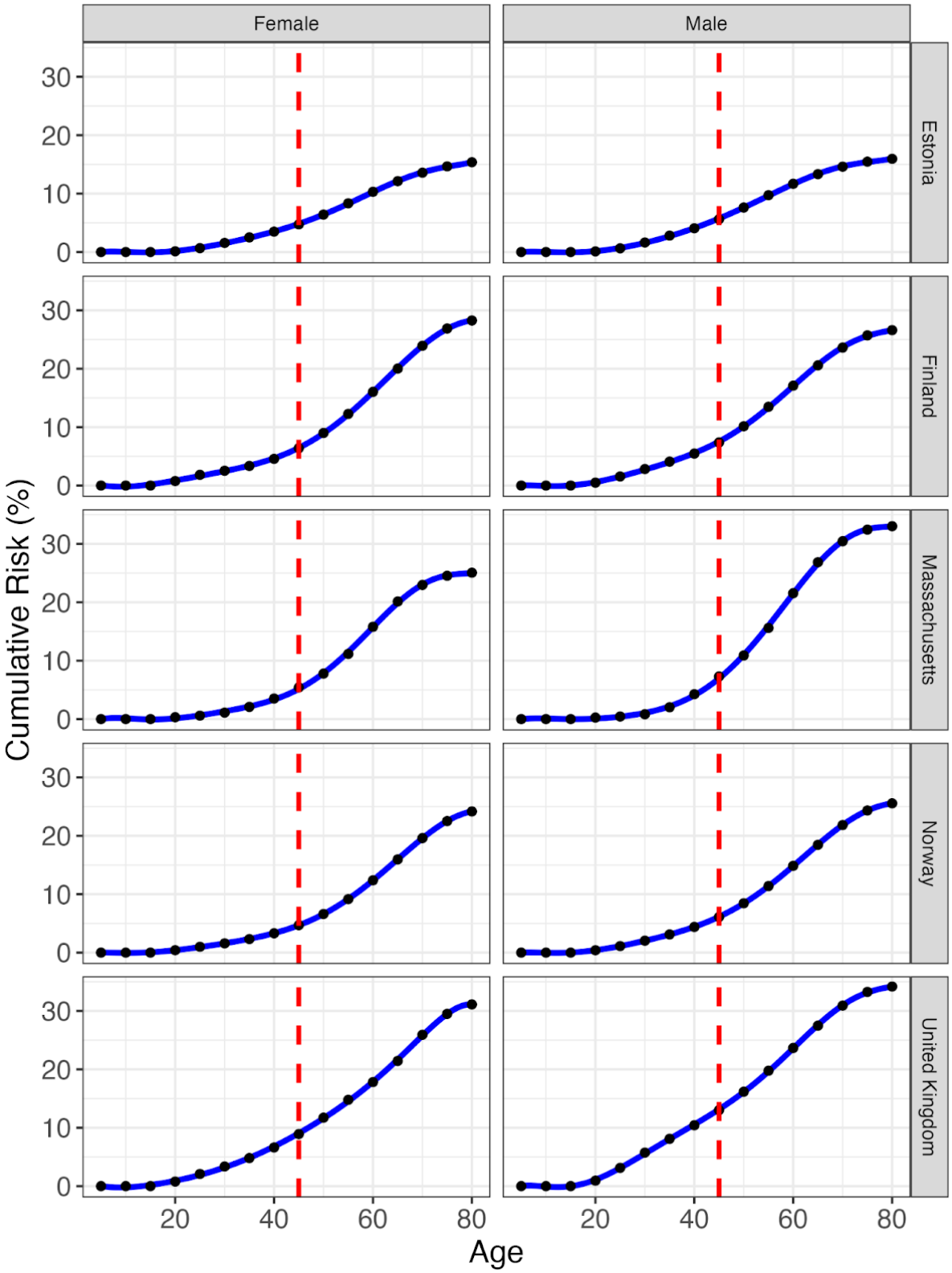


p. Type 2 Diabetes

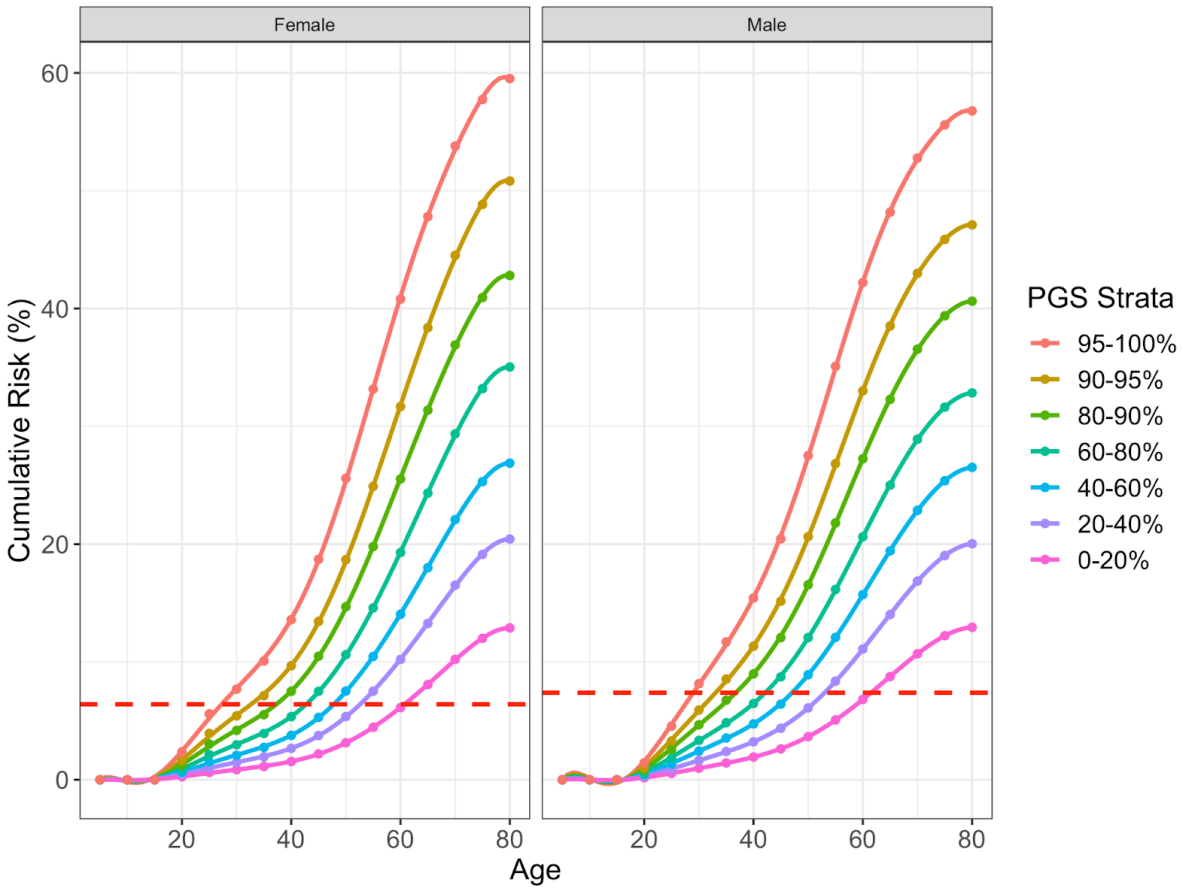


Supplementary Figure 6. Country and sex-specific cumulative absolute risk estimates in the top, bottom and reference percentile groupings also including uncertainty measures

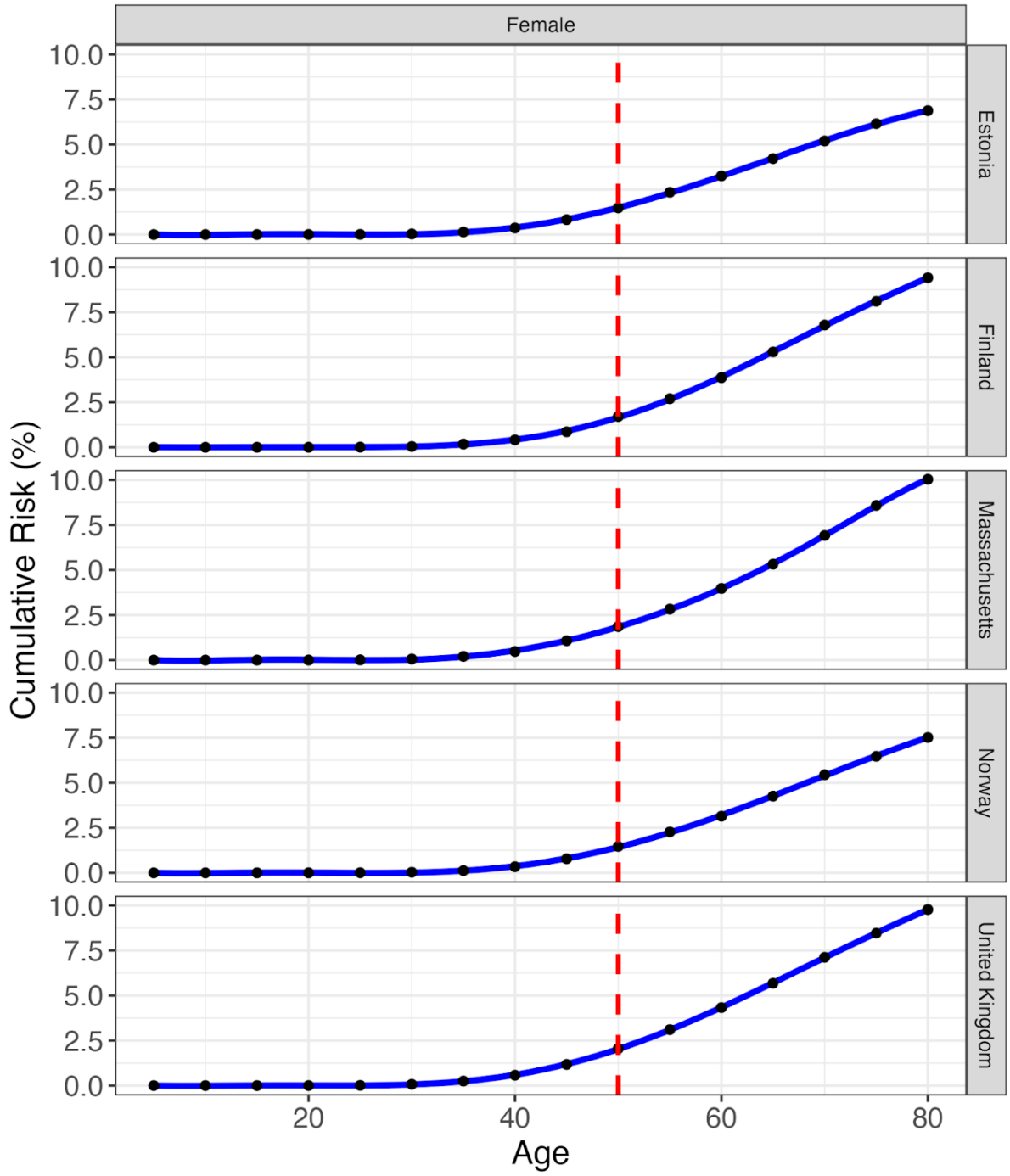
(95% confidence intervals). **a)** All Cancers. **b)** Appendicitis. **c)** Asthma. **d)** Atrial Fibrillation. **e)** Breast Cancer **f)** Colorectal Cancer. **g)** Epilepsy. **h)** Gout. **i)** Hip Osteoarthritis. **j)** Knee Osteoarthritis. **k)** Lung Cancer. **l)** Depression. **m)** Skin Melanoma. **n)** Rheumatoid Arthritis. **o)** Type 1 Diabetes. **p)** Type 2 Diabetes



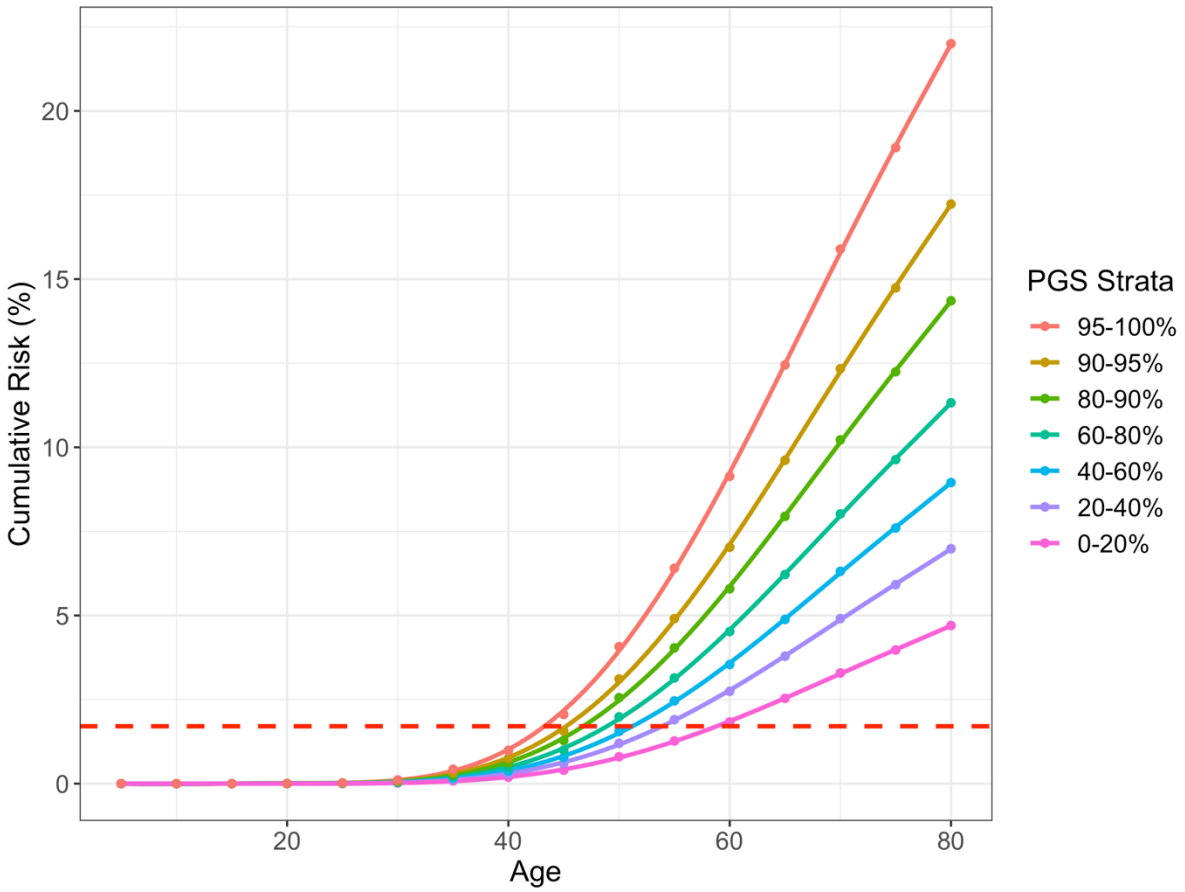
Supplementary Figure 7. Country-specific absolute risks for Type 2 Diabetes. Red dashed lines highlight age 45, the point at which the American Diabetes association recommends screening. The intersection with the blue cumulative risks were taken as the country's clinical threshold.



Supplementary Figure 8. Type 2 Diabetes cumulative incidence by PGS strata in the top and bottom percentile of the distribution and the reference category (40-60%) inclusive of confidence intervals.

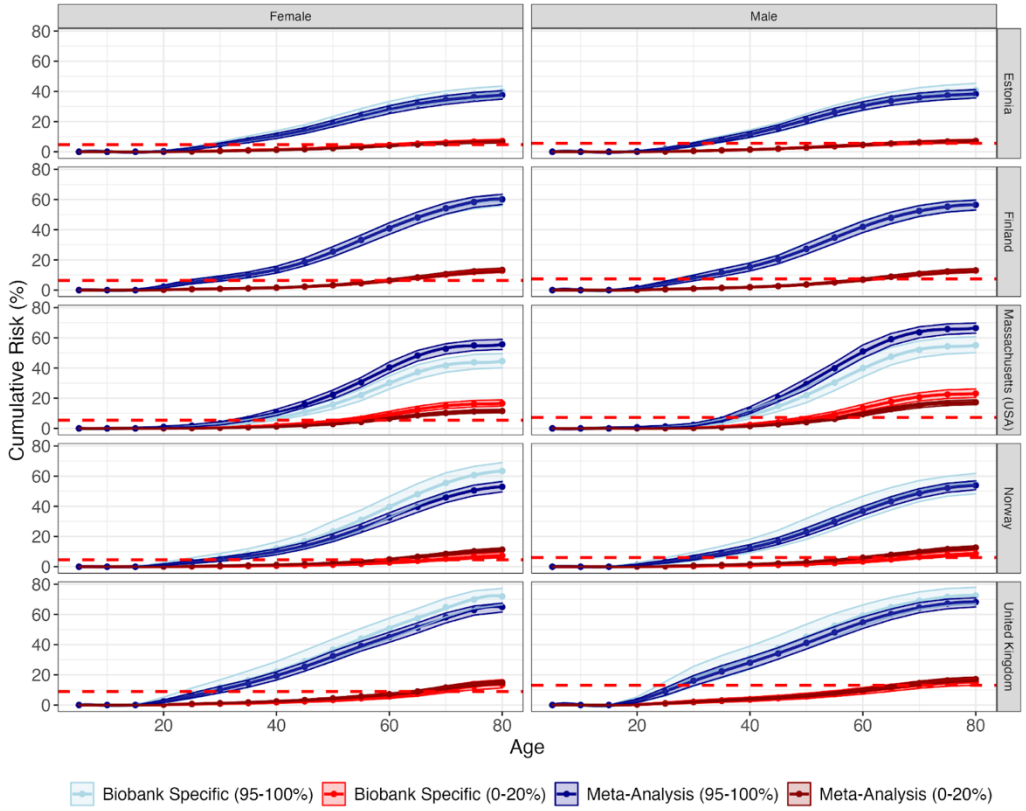


Supplementary Figure 9. Country-specific absolute risks for Breast Cancer. Red dashed lines highlight age 50, the age at which screening is recommended. The intersection with the blue cumulative risks were taken as the country's clinical threshold.

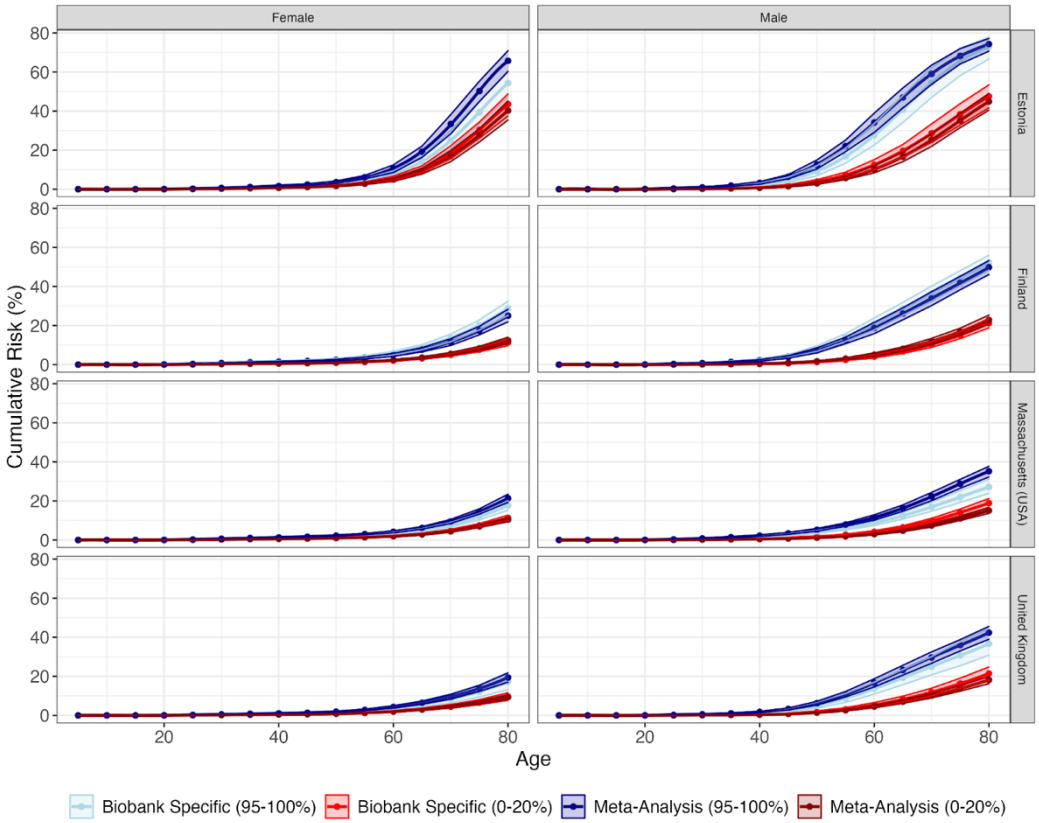


Supplementary Figure 10. Breast Cancer cumulative incidence by PGS strata in Finland inclusive of the top and bottom percentiles of risk for breast cancer.

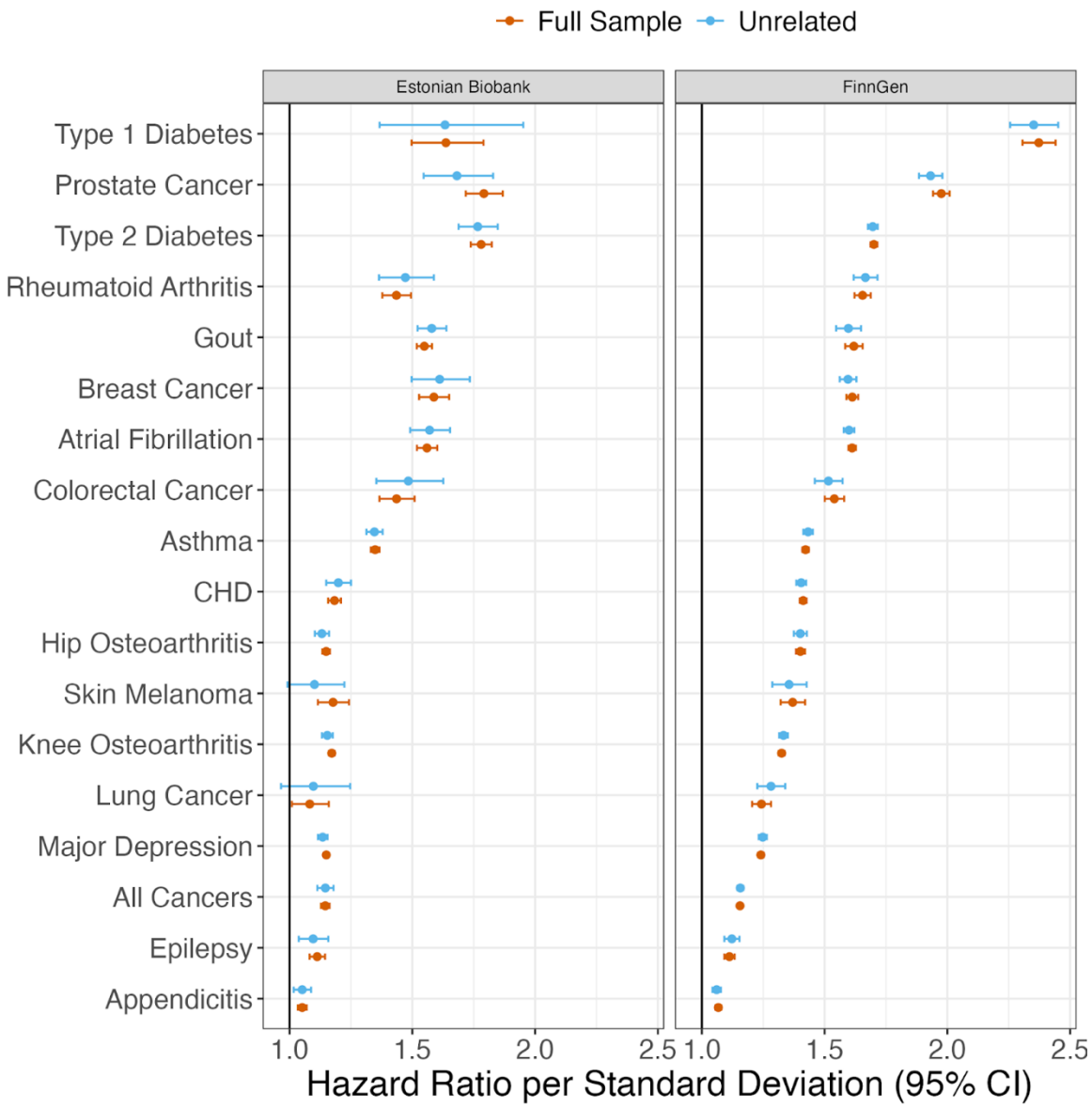
a)


Female
Male
Estonia
Finland
Massachusetts (USA)
Norway
United Kingdom
Cumulative Risk (%)
Age
Biobank Specific (95-100%)
Biobank Specific (0-20%)
Meta-Analysis (95-100%)
Meta-Analysis (0-20%)

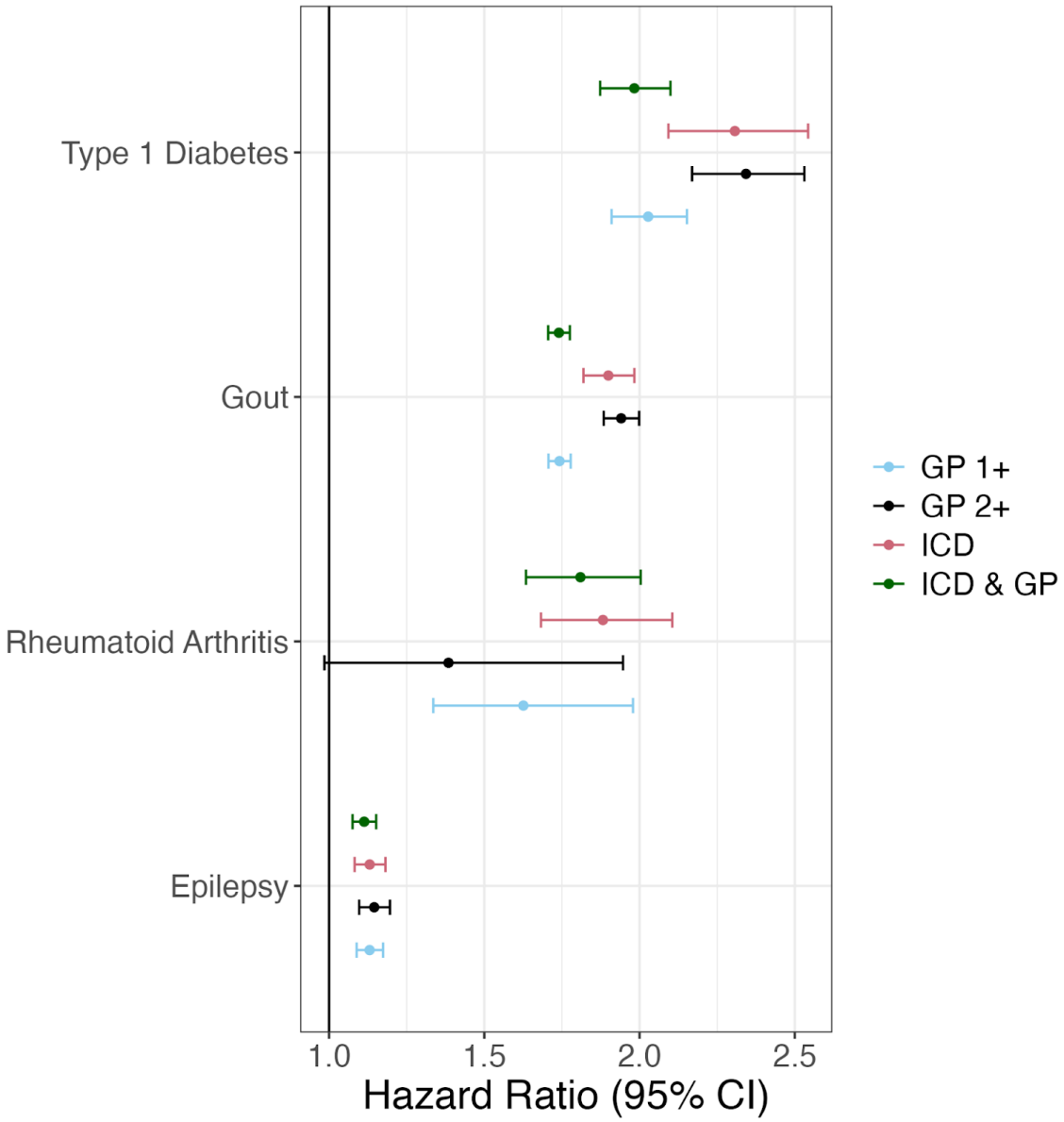
b)


Female
Male
Estonia
Finland
Massachusetts (USA)
United Kingdom
Cumulative Risk (%)
Age
Biobank Specific (95-100%)
Biobank Specific (0-20%)
Meta-Analysis (95-100%)
Meta-Analysis (0-20%)

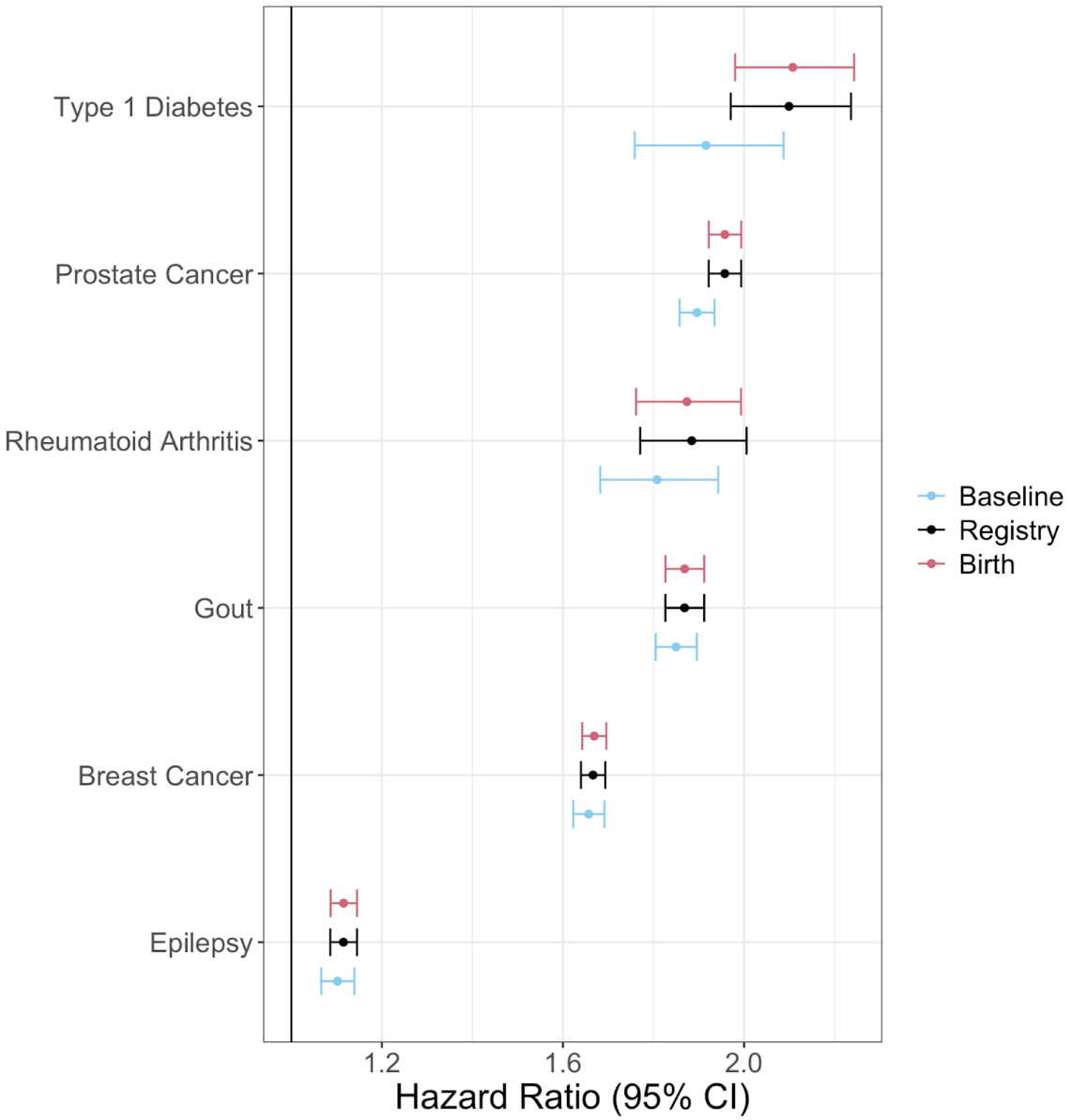
Supplementary Figure 11. Comparison of the cumulative incidence estimates resulting from the use of study specific hazard ratios or meta-analysed estimates. **a)** Type 2 Diabetes. **b)** Coronary Heart Disease



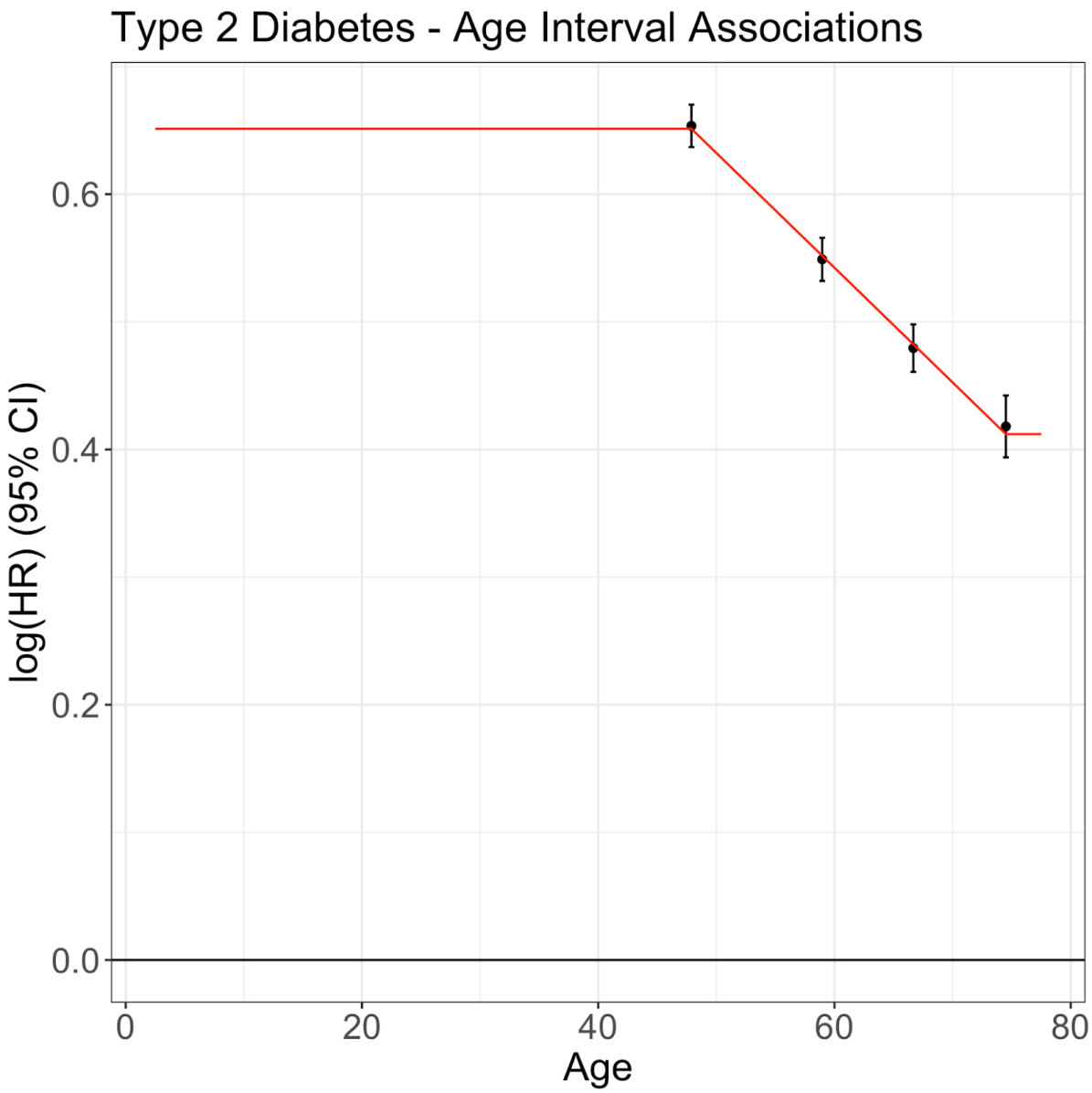
Supplementary Figure 12. Sensitivity analysis reviewing the impact of including relatives on the hazard ratios within FinnGen and Estonian Biobank. Third degree relatives and higher were removed and the hazard ratios were compared.



Supplementary Figure 13. Sensitivity analysis reviewing the impact of secondary care diagnoses (ICD codes) vs primary care diagnoses (Readv2 and CTV3 codes) on the hazard ratios within the UK Biobank.



Supplementary Figure 14. Sensitivity analysis reviewing the impact of assuming follow-up begins at birth, at the start of the registry linkage or at recruitment (baseline) in the UK Biobank.



Supplementary Figure 15. Estimating age-specific hazard ratios. This schematic describes how age-specific hazard ratios were estimated. Using Type 2 Diabetes and FinnGen data as an example, Cox Proportional Hazard models were first performed on four intervals. The log hazard ratios plotted above are placed at the median age at onset within each interval and a weighted linear regression was fit to the data. The predicted age specific log hazard ratios from the weighted linear regression are plotted in red. For any age outside of the range, the log hazard ratio is assumed to be equal to the nearest age in which a prediction was in range.

Supplementary Tables

Phenotype	PGS Beta (Std. Err)	PGS P-value	Sex Beta (Std. Err)	Sex P-value	PGS*Sex Beta (Std. Err)	PGS*Sex P-value
All Cancers	0.136 (0.004)	1.8x10-276	-0.007 (0.006)	0.2	-0.009 (0.006)	0.1
Appendicitis	0.053 (0.006)	1.27x10-18	0.036 (0.009)	1.6x10-4	0.018 (0.009)	0.05
Asthma	0.316 (0.004)	<2x10-16	-0.336 (0.008)	<2x10-16	0.036 (0.007)	6.71x10-7
Atrial Fibrillation	0.472 (0.007)	<2x10-16	0.641 (0.010)	<2x10-16	0.017 (0.009)	0.05
CHD	0.233 (0.007)	2.22x10-260	0.790 (0.010)	<2x10-16	0.063 (0.008)	1.2x10-13
Colorectal Cancer	0.345 (0.013)	1.88x10-153	0.275 (0.019)	<2x10-16	0.040 (0.018)	0.03
Epilepsy	0.110 (0.008)	6.7x10-39	0.138 (0.012)	<2x10-16	-0.019 (0.012)	0.11
Gout	0.443 (0.011)	<2x10-16	1.242 (0.015)	<2x10-16	0.069 (0.014)	3.56x10-7
Hip Osteoarthritis	0.216 (0.005)	<2x10-16	-0.247 (0.009)	<2x10-16	0.027 (0.009)	0.0024
Knee Osteoarthritis	0.218 (0.004)	<2x10-16	-0.265 (0.007)	<2x10-16	0.008 (0.007)	0.23
Major Depression	0.171 (0.003)	<2x10-16	-0.550 (0.007)	<2x10-16	0.009 (0.006)	0.16
Skin Melanoma	0.208 (0.013)	8.88x10-59	0.014 (0.020)	0.49	0.013 (0.019)	0.49
Rheumatoid Arthritis	0.493 (0.010)	<2x10-16	-0.807 (0.021)	<2x10-16	-0.004 (0.019)	0.82
Type 1 Diabetes	0.789 (0.016)	<2x10-16	0.274 (0.032)	<2x10-16	-0.020 (0.023)	0.38
Type 2 Diabetes	0.561 (0.005)	<2x10-16	0.316 (0.008)	<2x10-16	-0.054 (0.007)	4.78x10-14
Lung Cancer	0.191 (0.018)	5.33x10-26	0.393 (0.025)	<2x10-16	0.001 (0.024)	0.98

Supplementary Table 1. Main and interaction effects of polygenic scores and sex per phenotype. P-values are from a 2-sided Wald test.

Phenotype	Sex	Beta	SE	Pval	Delta	SE_Diff	Pval
All Cancers	male	-0.001	0.0004	0.04			
All Cancers	female	-0.0007	0.0004	0.15	-0.0002	0.0006	0.67
Appendicitis	male	0.0018	0.0005	0.02			
Appendicitis	female	0.0003	0.0002	0.12	0.0014	0.0006	0.01
Asthma	male	-0.0046	0.0005	0			
Asthma	female	-0.004	0.0003	0	-0.0006	0.0006	0.31
Atrial Fibrillation	male	-0.009	0.0006	0			
Atrial Fibrillation	female	-0.0053	0.0008	0	-0.0037	0.001	1.73E-04
CHD	male	-0.0079	0.0008	0			
CHD	female	-0.0013	0.0009	0.17	-0.0066	0.0012	1.43E-08
Colorectal Cancer	male	-0.0002	0.001	0.83			
Colorectal Cancer	female	-0.0032	0.0007	0	0.003	0.0012	0.01
Epilepsy	male	-0.002	0.0002	0			
Epilepsy	female	-0.0016	0.0008	0.08	-0.0004	0.0008	0.65
Gout	male	-0.0038	0.0004	0			
Gout	female	0.0014	0.0011	0.24	-0.0052	0.0011	4.51E-06
Hip Osteoarthritis	male	-0.0016	0.0002	0			
Hip Osteoarthritis	female	-0.0004	0.0006	0.53	-0.0013	0.0006	0.04
Knee Osteoarthritis	male	-0.0001	0.0002	0.71			
Knee Osteoarthritis	female	-0.0021	0.0003	0	0.0021	0.0004	3.67E-08
Major Depression	male	-0.001	0.0003	0.03			
Major Depression	female	-0.0009	0.0002	0	-0.0001	0.0004	0.89
Skin Melanoma	male	-0.003	0.0007	0.01			
Skin Melanoma	female	-0.0002	0.0004	0.66	-0.0028	0.0009	1.20E-03
Type 1 Diabetes	male	-0.0105	0.0011	0			
Type 1 Diabetes	female	-0.0093	0.0005	0	-0.0012	0.0012	0.34
Type 2 Diabetes	male	-0.0095	0.0006	0			
Type 2 Diabetes	female	-0.0063	0.0003	0	-0.0031	0.0007	4.16E-06
Lung Cancer	male	-0.0009	0.0003	0.03			
Lung Cancer	female	0.002	0.0012	0.14	-0.003	0.0012	0.02
Rheumatoid Arthritis	male	-0.0051	0.0006	0			
Rheumatoid Arthritis	female	-0.0055	0.0008	0	0.0004	0.0009	0.67

Supplementary Table 2. Differences across sex for betas resulting from weighting linear regression of log(Hazard Ratios) on Age**.** P-value (Pval) is calculated as a 2-sided Wald test.

No stratification	Sex stratified	Age stratified	Age and Sex stratified
Colorectal Cancer	Breast Cancer	All Cancers	Asthma
Lung Cancer	Coronary Heart Disease (Females)	Epilepsy	Atrial Fibrillation
Skin Melanoma	Hip Osteoarthritis	Major Depression	Coronary Heart Disease (Males)
Appendicitis (Females)	Gout (Females)	Rheumatoid Arthritis	Gout (Males)
Knee Osteoarthritis (Males)		Type 1 Diabetes	Knee Osteoarthritis (Females)
			Prostate Cancer
			Type 2 Diabetes
			Appendicitis (Males)

Supplementary Table 3. Disease-specific model selection. Note: Breast cancer and prostate cancer were only tested in females and males respectively. As such, by default, the analysis is sex stratified.

Country	Sex	Clinical Threshold
Estonia	Male	5.62
Estonia	Female	4.73
Finland	Male	7.38
Finland	Female	6.41
Massachusetts	Male	7.28
Massachusetts	Female	5.4
Norway	Male	6.1
Norway	Female	4.7
United Kingdom	Male	13.03
United Kingdom	Female	8.92

Supplementary Table 4. Country and sex-specific clinical thresholds for Type 2 Diabetes.

Country	Sex	Clinical Threshold
Estonia	Female	1.49
Finland	Female	1.71
Massachusetts	Female	1.86
Norway	Female	1.47
United Kingdom	Female	2.05

Supplementary Table 5. Country-specific clinical thresholds for Breast Cancer.

FollowUp Start Time	Phenotype	Controls	Cases	Beta	SE	Delta	SE_Diff	Z	Pval
Birth	Gout	439387	7945	0.63	0.01	NA	NA	NA	NA
Baseline	Gout	439377	6910	0.61	0.01	-0.01	0.02	-0.6	0.55
Registry	Gout	439387	7920	0.63	0.01	0	0.02	0	1
Birth	Prostate Cancer	193123	11638	0.67	0.01	NA	NA	NA	NA
Baseline	Prostate Cancer	193118	9457	0.64	0.01	-0.03	0.01	-2.3	0.02
Registry	Prostate Cancer	193123	11626	0.67	0.01	0	0.01	0	1
Birth	Rheumatoid Arthritis	446338	994	0.63	0.03	NA	NA	NA	NA
Baseline	Rheumatoid Arthritis	446328	733	0.59	0.04	-0.04	0.05	-0.74	0.46
Registry	Rheumatoid Arthritis	446338	978	0.63	0.03	0.01	0.04	0.13	0.9
Birth	Type 1 Diabetes	446340	992	0.75	0.03	NA	NA	NA	NA
Baseline	Type 1 Diabetes	446330	526	0.65	0.04	-0.1	0.05	-1.76	0.08
Registry	Type 1 Diabetes	446340	964	0.74	0.03	0	0.05	-0.09	0.93
Birth	Epilepsy	441446	5886	0.11	0.01	NA	NA	NA	NA
Baseline	Epilepsy	441436	3651	0.1	0.02	-0.01	0.02	-0.56	0.57
Registry	Epilepsy	441446	5736	0.11	0.01	0	0.02	-0.02	0.99
Birth	Breast Cancer	227339	15232	0.51	0.01	NA	NA	NA	NA
Baseline	Breast Cancer	227334	9055	0.5	0.01	-0.01	0.01	-0.55	0.58
Registry	Breast Cancer	227339	14954	0.51	0.01	0	0.01	-0.13	0.9

Supplementary Table 6. Differences in Hazard Ratio depending on if the start of follow-up assumed to begin at birth, registry linkage, or recruitment. P-value (Pval) is calculated as a 2-sided Wald test.

Phenotype	Global % of total Disability Adjusted Life Years	High-SDI Only % of total Disability Adjusted Life Years
All cancers	9.88	17.17
Colorectal cancer	0.96	2.07
Breast cancer	0.81	1.37
Type 2 diabetes	2.61	3.34
Prostate cancer	0.34	0.9
Coronary heart disease	7.19	7.42
Melanoma of skin	0.07	0.25
Asthma	0.85	1
Type 1 diabetes	0.18	0.25
Atrial fibrillation and flutter	0.33	0.84
Depression	1.46	1.94
Lung cancer	1.81	3.8
Seropositive rheumatoid arthritis	0.13	0.24
Hip-Osteoarthritis	0.04	0.13
Knee-Osteoarthritis	0.45	0.85
Gout	0.07	0.17
Epilepsy	0.52	0.43
Appendicitis	0.06	0.02
Total (exc. All cancers)	17.87	25.02

Supplementary Table 7. Diseases selected for analysis and contribution to global burden of disease as quantified by percentage of disability adjusted life years.

Supplementary References

1. Bellenguez, C. *et al.* A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinforma Oxf Engl* **28**, 134–135 (2012).
2. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
3. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955–959 (2012).
4. Mccarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016).
5. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
6. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
7. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).
8. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
9. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
10. Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443–1448 (2016).

11. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *Am J Hum Genet* **91**, 839–848 (2012).
12. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* **9**, 2643–2662 (2014).
13. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
14. Cavalli-Sforza, L. L. The Human Genome Diversity Project: past, present and future.Cavalli-Sforza LL. The Human Genome Diversity Project: past, present and future. *Nat Rev Genet* **6**, 333–340 (2005).
15. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
16. Castro, V. M. *et al.* The Mass General Brigham Biobank Portal: an i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. *J. Am. Med. Inform. Assoc.* **29**, 643–651 (2022).
17. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
18. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. Minimac2: Faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
19. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
20. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).

21. Oconnell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLOS Genet* **10**, (2014).
22. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
23. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
24. Mars, N. *et al.* Genome-wide risk prediction of common diseases across ancestries in one million people. *Cell Genom* **2**, None (2022).
25. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021).
26. Jia, G. *et al.* Evaluating the utility of polygenic risk scores in identifying high-risk individuals for eight common cancers. *JNCI Cancer Spectr.* **4**, kaa021 (2020).