**Subjects and Methods**

**Evaluation cohorts**

**MGI cohort.** Adult participants aged between 18 and 101 years at enrollment were recruited through the Michigan Medicine health system between 2012 and 2018 while awaiting diagnostic or interventional procedures either during a preoperative visit prior to the procedure or on the day of the procedure that required anesthesia. In addition to coded biosamples and secure, protected health information, participants understood that all EHR, claims, and national data sources linkable to the participant may be incorporated into the MGI databank. Each participant donated a blood sample for genetic analysis, underwent baseline vital sign testing, and completed a comprehensive history and physical assessment. We report results obtained from 38,360 unrelated, genotyped patients of inferred recent European ancestry with available integrated EHR data (~90 % of all MGI participants were inferred to be of recent European ancestry) 1. The data used in this study included diagnoses coded with the Ninth and Tenth Revision of the International Statistical Classification of Diseases (ICD9 and ICD10) with clinical modifications (ICD9-CM and ICD10-CM), sex, precomputed principal components (PCs), genotyping batch, and age. Data were collected according to the Declaration of Helsinki principles 2. MGI study participants’ consent forms and protocols were reviewed and approved by the University of Michigan Medical School Institutional Review Board (IRB ID HUM00099605 and HUM00155849). Opt-in written informed consent was obtained. Additional details about MGI can be found online (<https://precisionhealth.umich.edu/our-research/michigangenomics/>). A detailed comparison of the MGI versus UKB cohort (see below) can be found in Beesley et al, 2020 3.

**UK Biobank cohort (UKB).**UKB is a population-based cohort collected from multiple sites across the United Kingdom and includes over 500,000 participants aged between 40 and 69 years when recruited in 2006–2010 4. The open-access UK Biobank data used in this study included genotypes, ICD9 and ICD10 codes, inferred sex, inferred White British ancestry, kinship estimates down to third degree, birthyear, genotype array, and precomputed principal components of the genotypes.

**Genotyping, sample quality control and imputation**

*MGI*

DNA from 47,364 blood samples was genotyped on customized Illumina Infinium CoreExome-24 bead arrays and subjected to various quality control filters, resulting in a set of 392,323 polymorphic variants. Principal components and ancestry were estimated by projecting all genotyped samples into the space of the principal components of the Human Genome Diversity Project reference panel using PLINK (938 individuals) 5; 6. Pairwise kinship was assessed with the software KING 7, and the software FastIndep was used to reduce the data to a maximal subset that contained no pairs of individuals with 3rd-or closer degree relationship 8. We removed participants without EHR data and participants not of recent European descent from the analysis, resulting in a final sample of 38,360 unrelated subjects. Additional genotypes were obtained using the Haplotype Reference Consortium reference panel of the Michigan Imputation Server 9 and included over 24 million imputed variants with R2 ≥0.3 and minor allele frequency (MAF) ≥0.01%. Genotyping, quality control, and imputation are described in detail elsewhere 1.

*UK Biobank*

We used the UK Biobank Imputed Dataset (v3) and limited analyses to the documented 408,961 White British 10 individuals and 47,836,001 variants with imputation information score >= 0.3 and MAF >= 0.01% of which 22,846,729 overlapped with the imputed MGI data (see above). Two random subsets of 5,000 and 10,000 unrelated, White British individuals were used for LD analyses of UKB-based summary statistics.

**Phenome generation**

*MGI*

The MGI phenome was used as the discovery dataset and was based on ICD9-CM and ICD10-CM code data for 38,360 unrelated, genotyped individuals of recent European ancestry. Longitudinal time-stamped diagnoses were recoded to indicators for whether a patient ever had given a diagnosis code recorded by Michigan Medicine. These ICD9-CM and ICD10-CM codes were aggregated to form up to 1,857 PheCodes using the PheWAS R package (as described in detail elsewhere 1; 11). For each trait, we identified case and control samples by using the PheCode system where cases had at least one observed diagnosis code of the trait while controls (reference in fitted models) were individuals who did not have any diagnosis codes belonging to the trait and/or to the trait-specific PheCode exclusion list. To minimize differences in age and sex distributions, avoid extreme case-control ratios, and reduce the computational burden, we matched up to 10 controls to each case using the R package “MatchIt” 12. Nearest neighbor matching was applied for age and the first four principal components of the genotype data (PC1-4) using Mahalanobis distance with a caliper/width of 0.25 standard deviations. Exact matching was applied for sex and genotyping array. A total of 1,689 case-control studies with >50 cases were used for our analyses of the MGI phenome.

*UK Biobank*

The UK Biobank phenome was used as a replication dataset and was based on ICD9 and ICD10 code data of 408,961 White British 10, genotyped individuals that were similarly aggregated to PheCodes as MGI (as described elsewhere 13). In contrast to MGI, there were many pairwise relationships reported for UKB participants.

To retain a larger effective sample size for each phenotype, we first selected a maximal set of unrelated cases for each phenotype (defined as no pairwise relationship of 3rd degree or closer 8; 14) before selecting a maximal set of unrelated controls unrelated to these cases. Similar to MGI, we matched up to 10 controls to each case using the R package “MatchIt” 12. Nearest neighbor matching was applied for birthyear (as proxy for age, because age at diagnosis was not available to us) and PC1-4 (Mahalanobis-metric matching; matching window caliper/width of 0.25 standard deviations), and exact matching was applied for sex and genotyping array. A total of 1,419 case-control studies with >50 cases each were used for our analyses of the UK Biobank phenome.

On average, we were able to match 9 controls per case in the MGI phenome and 9.9 controls per case in the UKB phenome.

**PRS Structure**

PRS combine information across a defined set of genetic loci, incorporating each locus’s association with the target trait. The PRS for patient j takes the form PRS*j*= where *i* indexes the included loci for that trait, weight is the log odds ratios retrieved from the external GWAS summary statistics for locus *i,* and is a continuous version of the measured dosage data for the risk allele on locus *i* in subject *j.* In order to construct a PRS, one must determine which genetic loci to include in the PRS and their relative weights. Below, we obtain GWAS summary statistics from several different sources, resulting in several sets of weights for each trait of interest. For each set of weights, we consider several strategies for determining which genetic loci to include in the PRS construction.

**Sources of GWAS summary statistics**

For each of 68 cancers of interest, we collected GWAS summary statistics from up to three different sources: (1) merged genome-wide significant association signals published in the NHGRI EBI GWAS Catalog 15 if available; (2) large cancer GWAS meta-analysis if available; and (3) publicly available GWAS summary statistics of phenome x genome screening efforts of the UK Biobank data 13 (UKB GWAS Neale Lab, <http://www.nealelab.is/uk-biobank>; UKB GWAS Lee Lab, <https://www.leelabsg.org/resources>). If needed, we used LiftOver to convert coordinates of GWAS summary statistics to human genome assembly GRCh37 (UCSC Genome Browser Store, <https://genome-store.ucsc.edu>).

*GWAS Catalog*

We downloaded previously reported GWAS variants from the NHGRI-EBI GWAS Catalog (file version: r2019-05-03) 15; 16. Single nucleotide polymorphism (SNP) positions were converted to GRCh37 using variant IDs from dbSNP (build 151; UCSC Genome Browser) after updating outdated dbSNP IDs to their merged dbSNP IDs.

Entries with missing risk alleles, risk allele frequencies, or SNP-disease odds ratios were excluded. If a reported risk allele did not match any of the reported forward strand alleles of a non-ambiguous SNP (not A/T or C/G) in the imputed MGI genotype data (which correspond to the alleles of the imputation reference panel), we assumed minus-strand designation and corrected the effect allele to its complementary base of the forward strand. Entries with a reported risk allele that did not match any of the alleles of an ambiguous SNP (A/T and C/G) in our data were excluded at this step. We only included entries with broad European ancestry (as reported by the NHGRI-EBI GWAS Catalog) to match ancestries of discovery GWAS and target cohorts (MGI and UKB). As a quality control check, we compared the GWAS Catalog reported risk allele frequencies (RAF) with the RAF in MGI individuals. We then excluded entries whose RAF deviated more than 15%. This chosen threshold is subjective and was based on clear differentiation between correct and likely flipped alleles on the two diagonals, as noted frequently in GWAS meta-analyses quality control procedures 17.For SNPs with multiple entries, we kept the SNP with the most recent publication date (and smaller p-value, if necessary) and excluded the others.

*Large GWAS meta-analyses*

We downloaded full GWAS summary statistics made available by the “Breast Cancer Association Consortium” (BCAC) 14, the “Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome” (PRACTICAL) 18, and the “Ovarian Cancer Association Consortium” (OCAC) 19; 20. In addition, we extracted partial GWAS summary statistics that accompanied recent publications but were incomplete, i.e., reporting only SNPs below a certain p-value threshold 21-24. GWAS summary statistics were harmonized and, if needed, lifted over to human genome assembly GRCh37. In this paper, this source is referred to as “Large GWAS”.

*UK-Biobank-based GWAS*

We downloaded UK Biobank based GWAS summary statistics from two public repositories.

The first set of UK Biobank GWAS summary statistics were based on the analysis of up to 408,961 White British European-ancestry samples (UKB GWAS Lee Lab, <https://www.leelabsg.org/resources>). SNP-disease odds ratios were estimated using logistic mixed modeling adjusting for sample relatedness, and p-values were estimated using saddlepoint approximations (SAIGE method)25 to calibrate the distribution of score test statistics and, thus, control for unbalanced case-control ratios. The underlying phenotypes were auto-curated phenotypes based on the PheCodes of the PheWAS R package 1; 11; 13 similar to the phenomes used in our study and in the following are referred to as “UKB PHECODE”**.**

The second set of UK Biobank GWAS summary statistics were based on a linear regression model of up to 361,194 unrelated White British samples adjusting for relevant covariates (UKB GWAS Neale Lab, <http://www.nealelab.is/uk-biobank>). Three phenotype models were used in their analyses: (1) “PHESANT”: auto-curated phenotypes using PHEnome Scan ANalysis Tool, (2) “ICD10”: individuals with the same ICD10 category code (first three characters, e.g. “C50”) were used as cases while all non-coded individuals were treated as controls, and (3) “FINNGEN”: curated phenotypes / endpoints based on definitions of the Finngen consortium. In addition to the “UKB PHECODE” (described above), these three latter sources are referred to as “UKB PHESANT”, “UKB ICD10” and “UKB FINNGEN”, respectively (FINNGEN Clinical Endpoints, <https://www.finngen.fi/en/researchers/clinical-endpoints>; PHESANT, <https://github.com/MRCIEU/PHESANT>).

**PRS Construction**

For each set of GWAS summary statistics from the above-mentioned sources and each cancer, we develop up to seven different PRS using three different construction methods. Our goal of this approach was to compare multiple PRS methods and find the method that works best for the various types of GWAS summary statistics.

For the first two construction strategies, we performed LD clumping/pruning of variants with p-values below 10-4 by using the imputed allele dosages of 10,000 randomly selected samples and a pairwise correlation cut-off at r2 < 0.1 within 1Mb window. Using the resulting loci, we defined up to five sub-sets of variants with p-values below different thresholds (<5x10-9 to <5x10-5). These were used to construct a PRS tied to each threshold, where the PRS associated with p-values less than 5x10-8 is sometimes denoted as “GWAS hits.” For the second PRS construction method, we construct many different PRS across a fine grid of p-value thresholds. The p-value threshold with the highest cross-validated pseudo-R2 (see **PRS Evaluation** below) was used to define the more optimized “Pruning and Thresholding (P & T)” PRS.

As an alternative to the p-value thresholding and “P&T” PRS construction strategies, we also used the software package “lassosum” 26 and “PRS-CS” 27 to define a third type of PRS for GWAS sources with full summary statistics.  
Lassosum obtains PRS weights by applying elastic net penalization to GWAS summary statistics and incorporating LD information from a reference panel (<https://github.com/tshmak/lassosum>). Here, we used 5,000 randomly selected, unrelated samples as the LD reference panel. We applied a MAF filter of 1 % and, in contrast to the other two approaches, only included autosomal variants that overlap between summary statistics, LD reference panel, and target panel. Each “lassosum” run resulted in up to 76 combinations of the elastic net tuning parameters s and λ, and consequently, in 76 SNP sets with corresponding weights used to construct 76 PRS. We then selected the PRS with the highest pseudo-R2 to define the “lassosum” PRS (see **PRS Evaluation** below).

We used the software package “PRS-CS” 27 to define a PRS based on the continuous shrinkage (CS) priors (<https://github.com/getian107/PRScs>). PRS-CS uses a precomputed LD reference panel based on external European samples of the 1000 Genomes Project (“EUR reference”). We used the default settings (“PRS-CS-auto”), applied a MAF filter of 1 % and only included autosomal variants that overlap between summary statistics, LD reference panel, and target panel.

For each cancer and set of GWAS summary statistics, this approach resulted in up to eight PRS, where PRS with less than 5 included variants were excluded and the available GWAS summary statistics limited the available PRS construction techniques in some cases. Using the R package “Rprs” (<https://github.com/statgen/Rprs>), the value of each PRS was then calculated for each MGI participant and, if the GWAS source was not based on UKB, also for each UKB participant. For comparability of association effect sizes corresponding to the continuous PRS across cancer traits and PRS construction methods, we centered PRS values in MGI and UKB to their mean and scaled them to have a standard deviation of 1.

**PRS Evaluation**

For the PRS evaluations, except for when computing the pseudo-R2 (which is a measure of marginal association of the PRS with the outcome), we fit the following model for each PRS and cancer phenotype adjusting for covariates:

logit (P(Phenotype is present | PRS, Age, Sex, Array, PC)) =

We used Nagelkerke’s pseudo-R2 28 to select the tuning parameters within the “P&T” and lassosum construction methods (P-value for “P&T” SNP sets; s and λ for lassosum) and kept the PRS with the highest pseudo-R2 for further analyses. For each PRS derived for each GWAS source/method combination, we assessed the following performance measures relative to observed disease status in MGI and UKB:

(1) overall performance with Nagelkerke’s pseudo-R2 using R packages “rcompanion” 28, (2) accuracy with Brier score using R package “DescTools” 29; and (3) ability to discriminate between cases and controls as measured by the area under the covariate-adjusted receiver operating characteristic (AROC; semiparametric frequentist inference 30) curve (denoted AAUC) using R package “ROCnReg” 31.

For cross-validation purposes, we split the data corresponding to each trait in a phenome into training and test set. To retain case-control matching (see “Phenome generation” above), we randomly and equally distributed unique strata from matching and thus obtained a 50/50% split of cases where their matched controls were assigned to the same subset. We used the training set to determine the PRS tuning parameter(s) with the highest pseudo-R2 and used the testing set to obtain performance metric for that PRS. Firth's bias reduction method was used to resolve the problem of separation in logistic regression (R package “brglm2”)32; 33.

**PRS Association Testing**

Next, we assessed the strength of the relationship between these PRS and the traits they were designed for. To do this we fit the following model for each PRS and cancer phenotype adjusting for various covariates:

logit (P(Phenotype is present | PRS, Age, Sex, Array, PC)) =

where the PCs were the first four principal components obtained from the principal component analysis of the genotyped GWAS markers, where “Age” was the age at last observed diagnosis in MGI and birthyear in UKB and where “Array” represents the genotyping array. Our primary interest is , while the other factors (Age, Sex and PC) were included to address potential residual confounding and do not provide interpretable estimates due to the preceding application of case-control matching. Firth's bias reduction method was used to resolve the problem of separation in logistic regression (Logistf in R package “EHR”) 34-36.

To study the ability of the PRS to identify high risk patients, we fit the above model but replacing the PRS with an indicator for whether the PRS value was in the top 1, 2, 5, 10, or 25% (defined in controls) among the matched case control cohort.

**Phenome-wide Exploration of PRS Associations**

We selected PRS that were strongly associated with the cancer trait they were designed for phenome-wide association exploration in the phenomes of MGI and UKB for (p-value ≤ (0.05 / [#phenotypes in corresponding phenome]); see below).

We conducted PheWAS in MGI and also UKB (if the GWAS source was not based on UKB) to identify additional, secondary phenotypes associated with the PRS 37. To evaluate PRS-phenotype associations, we conducted Firth bias-corrected logistic regression by fitting model of equation 1 above for each PRS and each phenotype of the corresponding phenome. To adjust for multiple testing, we applied the conservative phenome-wide Bonferroni correction according to the total number of analyzed PheCodes (MGI: 1,689 phenotypes; UKB: 1,419 phenotypes). In Manhattan plots, we present –log10 (*p*-value) corresponding to tests of . Directional triangles on the PheWAS plot indicate whether a phenome-wide significant trait was positively (pointing up) or negatively (pointing down) associated with the PRS.

To investigate the possibility of the secondary trait associations with PRS being completely driven by the primary trait association, we performed a second set of PheWAS after excluding individuals affected with the primary or related cancer traits for which the PRS was constructed, referred to as “Exclusion-PRS-PheWAS” as described previously 1.

**Online Visual Catalog: *PRSweb***

The online open access visual catalog *PRSweb* was implemented using Grails, a Groovy- and Java-based backend logic, to integrate interactive visualizations and MySQL databases. Interactive PheWAS plots are drawn with the JavaScript library “LocusZoom.js” which is maintained by the UM Center for Statistical Genetics (Locuszoom, <https://github.com/statgen/locuszoom>) and offers dynamic plotting, automatic plot sizing, and label positioning. Additional data-driven visualizations (e.g., temporal relationship plots) were implemented with the JavaScript library “D3.js”.

Unless otherwise stated, analyses were performed using R 3.6.1 38

**References**

1. Fritsche, L.G., Gruber, S.B., Wu, Z., Schmidt, E.M., Zawistowski, M., Moser, S.E., Blanc, V.M., Brummett, C.M., Kheterpal, S., Abecasis, G.R., et al. (2018). Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. Am J Hum Genet 102, 1048-1061.

2. World Medical Association. (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 310, 2191-2194.

3. Beesley, L.J., Salvatore, M., Fritsche, L.G., Pandit, A., Rao, A., Brummett, C., Willer, C.J., Lisabeth, L.D., and Mukherjee, B. (2020). The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. Stat Med 39, 773-800.

4. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12, e1001779.

5. Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H.M., Stambolian, D., Chew, E.Y., Branham, K.E., Heckenlively, J., Study, F., Fulton, R., et al. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. Nat Genet 46, 409-415.

6. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100-1104.

7. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867-2873.

8. Abraham, K.J., and Diaz, C. (2014). Identifying large sets of unrelated individuals and unrelated markers. Source Code Biol Med 9, 6.

9. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48, 1279-1283.

10. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv.

11. Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 30, 2375-2376.

12. Ho, D.E., Imai, K., King, G., and Stuart, E.A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. J Stat Softw 42, 1-28.

13. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet 50, 1335-1341.

14. Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemacon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92-94.

15. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45, D896-D901.

16. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42, D1001-1006.

17. Winkler, T.W., Day, F.R., Croteau-Chonka, D.C., Wood, A.R., Locke, A.E., Magi, R., Ferreira, T., Fall, T., Graff, M., Justice, A.E., et al. (2014). Quality control and conduct of genome-wide association meta-analyses. Nat Protoc 9, 1192-1212.

18. Schumacher, F.R., Al Olama, A.A., Berndt, S.I., Benlloch, S., Ahmed, M., Saunders, E.J., Dadaev, T., Leongamornlert, D., Anokian, E., Cieza-Borrella, C., et al. (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet 50, 928-936.

19. Phelan, C.M., Kuchenbaecker, K.B., Tyrer, J.P., Kar, S.P., Lawrenson, K., Winham, S.J., Dennis, J., Pirie, A., Riggan, M.J., Chornokur, G., et al. (2017). Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. Nat Genet 49, 680-691.

20. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47, D1005-D1012.

21. Ransohoff, K.J., Wu, W., Cho, H.G., Chahal, H.C., Lin, Y., Dai, H.J., Amos, C.I., Lee, J.E., Tang, J.Y., Hinds, D.A., et al. (2017). Two-stage genome-wide association study identifies a novel susceptibility locus associated with melanoma. Oncotarget 8, 17586-17592.

22. Huyghe, J.R., Bien, S.A., Harrison, T.A., Kang, H.M., Chen, S., Schmit, S.L., Conti, D.V., Qu, C., Jeon, J., Edlund, C.K., et al. (2019). Discovery of common and rare genetic risk variants for colorectal cancer. Nat Genet 51, 76-87.

23. Chahal, H.S., Wu, W., Ransohoff, K.J., Yang, L., Hedlin, H., Desai, M., Lin, Y., Dai, H.J., Qureshi, A.A., Li, W.Q., et al. (2016). Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. Nat Commun 7, 12510.

24. Chahal, H.S., Lin, Y., Ransohoff, K.J., Hinds, D.A., Wu, W., Dai, H.J., Qureshi, A.A., Li, W.Q., Kraft, P., Tang, J.Y., et al. (2016). Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. Nat Commun 7, 12048.

25. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet 50, 1335-1341.

26. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet Epidemiol 41, 469-480.

27. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun 10, 1776.

28. Mangiafico, S. (2019). rcompanion: Functions to Support Extension Education Program Evaluation. In. (

29. Signorell, A. (2018). DescTools: Tools for Descriptive Statistics. In. (

30. Janes, H., and Pepe, M.S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. Biometrika 96, 371-382.

31. Rodriguez-Alvarez, M.X., and Inacio, V. (2020). ROCnReg: ROC Curve Inference with and without Covariates. In. (

32. Kosmidis, I., Clovis Kenne Pagui, E., and Sartori, N. (2018). Mean and median bias reduction in generalized linear models. In arXiv e-prints. p arXiv:1804.04085.

33. Kosmidis, I. (2019). brglm2: Bias Reduction in Generalized Linear Models. In. (

34. Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. Stat Med 25, 4216-4226.

35. Heinze, G., Ploner, M., Dunkler, D., and Southworth, H. (2013). logistf: Firth's bias reduced logistic regression. In. (

36. Choi, L., and Beck, C. (2017). EHR: Electronic Health Record (EHR) Data Processing and Analysis Tool. In. (

37. Fritsche, L.G., Gruber, S.B., Wu, Z., Schmidt, E.M., Zawistowski, M., Moser, S.E., Blanc, V.M., Brummett, C.M., Kheterpal, S., Abecasis, G.R., et al. (2018). Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. Am J Hum Genet 102, 1048-1061.

38. R Core Team. (2016). R: A Language and Environment for Statistical Computing. In. (R Foundation for Statistical Computing, Vienna, Austria.