Supplementary Information

PGS Browser: a public platform for personalized polygenic score analysis and interpretation

Nikita Kolosov[1-3], Mary P. Reeve[3,4], Pietro Della Briotta Parolo[3], Mitja I. Kurki[3-5], FinnGen[#], Vincent Llorens[3], Timo Petteri Sipila[3], Adam Herman[1], Ivan Molotkov[1-3], Mervi Aavikko[3], Samuli Ripatti[3], Aarno Palotie[3,4], Mark J. Daly[3-5], Mykyta Artomov[1-3*]

[1] – The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA

[2] – Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, USA

[3] – Institute of Molecular Medicine Finland (FIMM), Helsinki Institute of Life Sciences (HiLIFE), University of Helsinki, Helsinki, Finland

[4] – Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

[5] – Analytic and Translational Genetics Unit, Massachusetts General Hospital (MGH), Boston, MA, USA

[#] - a list of FinnGen authors and their affiliations available at the end of the paper, email: finngen-info@helsinki.fi

[*] - correspondence: mykyta.artomov@nationwidechildrens.org

Table of contents:

Ethics Statement

Ethics statement and materials & methods:
Study subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017. The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and THL/1524/5.05.00/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020), Findata permit numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020, THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020, THL/6619/14.06.00/2020, THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021, THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/2021, Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20) TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and Finnish Registry for Kidney Diseases permission/extract from the meeting minutes on 4th July 2019. The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 11 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, BB2021_65, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, HUS/248/2020, HUS/430/2021 §28, §29, HUS/150/2022 §12, §13, §14, §15, §16, §17, §18, §23, §58 and §59, Auria Biobank AB17-5154 and amendment #1 (August 17 2020) and amendments BB_2021-0140, BB_2021-0156 (August 26 2021, Feb 2 2022), BB_2021-0169, BB_2021-0179, BB_2021-0161, AB20-5926 and amendment #1 (April 23 2020) and it´s modification (Sep 22 2021), BB_2022-0262, BB_2022-0256, Biobank Borealis of Northern Finland_2017_1013, 2021_5010, 2021_5018, 2021_5015, 2021_5015 Amendment, 2021_5023, 2021_5023 Amendment, 2021_5017, 2022_6001, 2022_6006 Amendment, BB22-0067, 2022_0262, Biobank of Eastern Finland 1186/2018 and amendment 22§/2020, 53§/2021, 13§/2022, 14§/2022, 15§/2022, 27§/2022, 28§/2022, 29§/2022, 33§/2022, 35§/2022, 36§/2022, 37§/2022, 39§/2022, 7§/2023, Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020), 8§/2021, 9§/2021, §9/2022, §10/2022, §12/2022, 13§/2022, §20/2022, §21/2022, §22/2022, §23/2022, 28§/2022, 29§/2022, 30§/2022, 31§/2022, 32§/2022, 38§/2022, 40§/2022, 42§/2022, 1§/2023, Central Finland Biobank 1-2017, BB_2021-0161, BB_2021-0169, BB_2021-0179, BB_2021-0170, BB_2022-0256, and Terveystalo Biobank STB 2018001 and amendment 25th Aug 2020, Finnish Hematological Registry and Clinical Biobank decision 18th June 2021, Arctic biobank P0844: ARC_2021_1001.

PGS models processing

Most preprocessing steps were carried out using the pgscatalog-utils Python package (see Web Resources). We began by downloading 3,688 PGS Catalog models using the download_scorefiles function. For computational efficiency and parallel processing, all models were then merged into five separate matrices using combine_scorefiles. To harmonize the PGS models (GRCh38) with FinnGen genotype data (Release 11), we used the .bim file and applied the match_variants function, using the default matching threshold of 75%. Models that did not meet this threshold were excluded from further analysis. PGS scores were computed for each pair of matching score files (*_ALL_additive_[01].scorefile.gz.tsv) using PLINK's --score function with the cols=+scoresums,+denom arguments. The final score for each individual was calculated as the sum of matched SNP weights divided by the number of non-missing alleles. Finally, we manually reviewed each model for potential technical inconsistencies, including misreported effect alleles, incorrect trait labels, and population mismatches between the GWAS cohort and FinnGen.

PGS evaluation measures

Below we provide a detailed explanation of the evaluation measures used in an article along with the guidance on how these measures should be interpreted in practice (Supplementary Table 1). Importantly, most of these metrics are not intended to define clinical decision thresholds (e.g. 90%) in the current study. Rather, they are reported to characterize the statistical properties of PGS models beyond AUROC and to facilitate standardized comparison across models and phenotypes. PGS Browser provides an interactive interface in which some of these threshold dependent metrics can be explored across the full range of percentile thresholds (see https://pgs.nchigm.org/, Classification measures tab), allowing users to select cutoff appropriate for their specific research or clinical context. As an example, Supplementary Figure 8 provides some of these metrics for top 15 PGSs.

1) Unadjusted area under the receiver operating characteristic curve (ROC $AUC_{unadjust.}$)

Practical interpretation: ROC AUC reflects the overall ability of a PGS model to discriminate between cases and controls across all possible thresholds, and ranges from 0 to 1. Higher values (closer to 1) indicate stronger genetic risk stratification. However, ROC AUC does not directly translate into individual risk estimates or clinical decision thresholds and therefore should be interpreted primarily as a measure of comparative predictive performance between models rather than a metric of clinical actionability.

ROC AUC and corresponding confidence intervals (CIs) were calculated using AUCBoot function from the bigsnpr R package [1]. It is formally defined as the probability that the predictive model will rank a randomly chosen positive instance higher than a randomly chosen negative instance as follows:

$$\mathrm{ROC\ AUC} = \mathrm{P}\left(S(x_{\mathrm{pos}}) > S(x_{\mathrm{neg}})\right), \quad (1)$$

where $S(x)$ is the score assigned by the model to instance x. x_{pos} and x_{neg} are randomly chosen positive and negative instances, respectively. Note that sometimes ROC AUC can be less than 0.5, which means that cases are placed by PGS model lower than controls. In most of the cases this is the result of effective allele misreporting, the result of the extremely small number of cases (n<5) or drastic ancestry mismatch between target cohort and GWAS used to derive the PGS model. Note, that ROC AUC using this method was calculated for raw PGS vectors, without including any covariates.

2) Variance explained on liability scale

Practical interpretation: Variance explained on the liability scale (R^2) quantifies the proportion of genetic liability captured by the PGS under the liability threshold model, and ranges from 0 to 1 [2]. Higher values (closer to 1) indicate stronger contribution of the score to disease risk prediction. This metric is best interpreted as a measure of the overall genetic contribution to prediction rather than as a direct indicator of clinical utility.

In this article we reported incremental value, calculated as a difference between full model value (with covariates PGS, Age, PC1-6) and null model (excluding PGS). In order to calculate this variance we used the approach described in Lee et al [2] (and here). As a source of population prevalence, we used FinnRegistry. If the endpoint lacked prevalence information in FinnRegistry we used prevalence from the newer FinnGen R13 release (see Supplementary Data 4). Below we provide code used to calculate R^2 on liability scale:

```
# R2 on the liability scale using the transformation
nt = total number of the sample
ncase = number of cases
ncont = number of controls
thd = the threshold on the normal distribution which truncates the proportion of disease prevalence
K = population prevalence
P = proportion of cases in the case-control samples
#threshold
thd = -qnorm(K,0,1)
#value of standard normal density function at thd
zv = dnorm(thd)
#mean liability for case
mv = zv/K
#linear model
lmv = lm(y~g)
#R20 : R2 on the observed scale
R2O = var(lmv$fitted.values)/(ncase/nt*ncont/nt)
# calculate correction factors
theta = mv*(P-K)/(1-K)*(mv*(P-K)/(1-K)-thd)
cv = K*(1-K)/zv^2*K*(1-K)/(P*(1-P))
# convert to R2 on the liability scale
R2 = R2O*cv/(1+R2O*theta*cv)
```

3) Odds Ratio/Hazard Ratio

Practical interpretation: Odds ratios (OR) and hazard ratios (HR) quantify the strength of association between the PGS and disease risk per standard deviation increase in the score. Larger values (>1) indicate stronger genetic risk. However, these values should not be interpreted

in isolation when comparing across diseases. For example, an OR of 4 for vitiligo reflects stronger genetic discrimination than an OR of 1.5 for thyroid disorders, but the clinical relevance of these values depends on additional factors such as disease prevalence or disease severity.

Odds Ratio was estimated using logistic regression. HR was estimated using a cox-proportional hazard model. Both were calculated as an exponent of the β_{PGS} coefficient produced by the selected models. We adjusted the relationship between standardized PGS and each endpoint for Sex, Age, PC1-6, and Genotyping Array (see Materials and Methods, Phenome-Wide Association Study).

4) Recall (True Positive Rate) and Specificity (True Negative Rate)

Practical interpretation: Recall (or TPR) indicates the fraction of true cases captured at a given PGS threshold, while specificity (or TNR) reflects the fraction of controls correctly excluded. These measures illustrate the trade-off between identifying more true cases and limiting false positives. The optimal balance between recall and specificity depends on the intended application of the model, such as screening, risk stratification, or research recruitment.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{3}$$

, where TP = true positives, FN = false negatives, TN = true negatives, FP = false positives.

5) Positive Predictive Value (PPV), False Discovery Rate (FDR) and Standardized Positive Predictive Value (PPV_{std})

Practical interpretation: The Positive Predictive Value (PPV) or precision represents the probability that individuals exceeding a selected PGS threshold truly have the disease. Because PPV strongly depends on disease prevalence, we report standardized PPV (PPV_{std})[3] assuming a reference prevalence of 5% to facilitate comparisons across phenotypes. For example, a PPV of 0.1 under a standardized prevalence of 0.05 corresponds to approximately a two-fold enrichment of risk relative to baseline prevalence. Such enrichment may support targeted screening, monitoring, or research recruitment.

Opposite to this PPV is False Discovery Rate (FDR):

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}} = 1 - \mathrm{PPV} \tag{5}$$

During PPV standardization specificity and sensitivity get fixed and reference prevalence gets selected. We choose a prevalence of 5% as a reference value, since it represents around the median value of prevalence of diseases tested. Adjustment procedure goes like that:

$$\mathrm{PPV}_{\mathrm{std}} = \frac{\mathrm{Sensitivity} \cdot \mathrm{Prevalence}}{(\mathrm{Sensitivity} \cdot \mathrm{Prevalence}) + \big((1 - \mathrm{Specificity}) \cdot (1 - \mathrm{Prevalence})\big)} \quad (6)$$

6) Absolute/Relative Risk

Practical interpretation: Relative risk quantifies the degree of risk enrichment among individuals exceeding a specified PGS threshold compared with a reference group (e.g., the median PGS). Absolute risk estimates the probability of disease conditional on the selected threshold and disease prevalence in the analyzed cohort. These measures provide complementary perspectives: relative risk reflects the strength of genetic stratification, while absolute risk provides a more intuitive estimate of disease probability for individuals within a given risk stratum.

Absolute risk was estimated via Bayes' theorem as the posterior probability of having disease given a specific PGS threshold:

$$\Pr(\,\mathrm{Case} \mid \mathrm{PGS}\,) = \frac{\Pr(\,\mathrm{PGS} \mid \mathrm{Case}\,) \cdot \Pr(\mathrm{Case})}{\Pr(\,\mathrm{PGS} \mid \mathrm{Case}\,) \cdot \Pr(\mathrm{Case}) + \Pr(\,\mathrm{PGS} \mid \mathrm{Control}\,) \cdot \Pr(\mathrm{Control})} \quad (7)$$

Relative Risk was defined as the probability of being a case for the selected PGS threshold divided by the probability of being the case at 50th PGS percentile.

7) Time-dependent area under the ROC curve ($_{td}$ROC AUC)

Practical interpretation: The $_{td}$ROC AUC reflects how well a survival model can distinguish between individuals who experience an event earlier versus later over the specified time horizon (e.g. 1-15 years), and ranges between 0 and 1. Higher $_{td}$ROC AUC values indicate stronger ability of the model to rank individuals according to their relative risk of developing the outcome over time. However, this measure captures only discrimination and does not assess whether the predicted survival probabilities are well calibrated. Therefore, $_{td}$ROC AUC is most useful for comparing the relative predictive performance of different survival models rather than evaluating the absolute accuracy of predicted risk.

It was calculated using cumulative_dynamic_auc function from scikit-survival (v.0.23.0) Python package. Selected time-points are mentioned in corresponding figure captions. In order to estimate the confidence intervals, we bootstrapped the test set, calculated average $_{td}$ROC AUC across different time points and, in the end, derived CIs from the sample of 100 average $_{td}$ROC AUCs.

8) D-calibration

Practical interpretation: D-calibration [4] evaluates whether predicted survival probabilities are consistent with the observed event frequencies across the population. A non-significant p-value ($p > 0.05$) indicates that predicted risks are statistically consistent with observed outcomes and therefore that the survival model is well calibrated. In practical terms, this means that predicted event probabilities can be interpreted as reliable estimates of risk over time, which is particularly important when models are used to estimate absolute risk rather than simply rank individuals by relative risk.

P-value for D-calibration was calculated using SurvivalEval (v.0.3.0) python package.

Metric	What it measures	How it is calculated in this study	Interpretation	Threshold	Notes
ROC AUC	Overall discrimination ability of the PGS	Probability that a randomly selected case has a higher PGS than a randomly selected control	0.5 = no discrimination; 1.0 = perfect discrimination	Not threshold-dependent	Reported for raw PGS without covariates
Variance explained (R^2 liability)	Contribution of PGS to disease risk variance	Incremental R^2 between the full model (PGS + covariates) and the null model	Higher values indicate stronger genetic contribution	Not threshold-dependent	Converted to liability scale using population prevalence
Odds Ratio / Hazard Ratio	Strength of association between standardized PGS and outcome	Estimated from logistic or Cox regression models	OR/HR > 1 indicates higher risk with increasing PGS	Reported for 1 standard deviation increase in PGS	Adjusted for age, sex, PCs, and genotyping array
Recall (True Positive Rate)	Proportion of true cases captured by the selected PGS threshold	$TP/(TP + FN)$	Higher recall means more cases identified	90th percentile	Illustrates detection ability at a fixed risk threshold
Specificity (True Negative Rate)	Proportion of controls correctly excluded	$TN/(TN + FP)$	Higher specificity means fewer false positives	90th percentile	Complementary to recall
PPV (Precision)	Probability that individuals above threshold are true cases	$TP/(TP + FP)$	Higher PPV indicates stronger risk enrichment	90th percentile	Strongly dependent on disease prevalence
Standardized PPV (PPV_{std})	PPV adjusted for prevalence differences	PPV recalculated assuming a 5% reference prevalence	Allows comparison across diseases with different prevalence	90th percentile	Reference prevalence = 5%
Relative Risk	Risk enrichment in high-PGS group	Risk above threshold divided by risk at the 50th percentile	Values > 1 indicate increased disease risk in the high-PGS group	90th vs 50th percentile	Used for model comparison
Absolute Risk	Estimated probability of disease given a specific PGS threshold	Derived using Bayes' theorem	Reflects expected disease probability in the high-PGS group	90th percentile	Strongly dependent on disease prevalence
Time-dependent ROC AUC	Discrimination for survival models	ROC AUC calculated across selected time points and averaged	Same interpretation as ROC AUC	Not threshold-dependent	Used for time-to-event prediction models
Integrated Brier Score (IBS)	Overall survival model performance	Mean squared prediction error over time	Lower values indicate better performance	Not threshold-dependent	Captures discrimination and calibration
D-calibration	Calibration of survival predictions	Goodness-of-fit test comparing predicted and observed event probabilities	$p > 0.05$ indicates good calibration	Not threshold-dependent	Evaluates survival probability calibration

Supplementary Table 1. Interpretation of PGS evaluation metrics used in this study. Threshold-dependent metrics were calculated using the 90th percentile of the PGS distribution to enable standardized comparison across models and phenotypes.

Ancestry prediction in FinnGen cohort

To model continental population structure, we first selected 51,137 HapMap3 variants shared between the 1000 Genomes Project whole-genome sequencing (WGS) data and the FinnGen genotyping array. Restricting principal components analysis (PCA) to these variants we tried to exclude the batch effect introduced by the imputation and at the same time keep variants informative for the ancestry prediction in FinnGen samples. Using PLINK's -pca we derived six principal components and projected all 473,681 FinnGen samples into this 1000G PC space (Figure 6a). Each PC-score was standardized by dividing by the square root of its eigenvalue and scaling by –2. Further, we used these PCs to infer ancestry for non-Finnish samples.

To infer ancestry for non-Finnish samples, we trained an XGBoost classifier (xgboost, v3.0.2) on 1000 Genomes data (90% training, 10% testing). We optimized the following hyperparameters using grid search (cv=10) and one-vs-rest ROC AUC as a scoring function: n_estimators, max_depth, learning_rate, subsample, colsample_bytree, and reg_lambda. The final model achieved a weighted F1-score of 0.99 on the test set.

Using the tuned XGBoost classifier, we assigned continental ancestries to 19,947 non-Finnish FinnGen participants: 15,101 non-Finnish Europeans, 1,620 Admixed Americans, 690 East Asians, 433 Africans, and 352 South Asians. Individuals whose highest ancestry probability fell below 65% (n = 1,751) were labeled as Other.

Disease risk prediction models

To identify diseases that benefit from integrating multiple PGSs for diagnosis prediction or future incidence risk, we developed predictive models for both disease status classification and age-of-onset prediction. These models were designed to assess whether incorporating multiple polygenic scores improves predictive performance beyond the use of a single best-performing PGS. All models were trained and tested on 453,734 Finnish ancestry samples with complete covariate data. To make sure that there is no sample overlap between GWASs used for PGS models development and FinnGen samples, for further experiments we utilized solely top 6 publications, covered almost the half of all available models (n=1,514), which reliably lack any FinnGen sample overlap, naming: Tanigawa Y et al. (PGP000244), Privé F et al. (PGP000263), Weissbrod O et al. (PGP000332), Liu N et al. (PGP000464), Ding Y et al. (PGP000457) and Sinnott-Armstrong N et al. (PGP000128) [5–10].

For each disease endpoint, we trained three distinct models with different sets of covariates. First, null model - included sex, age, and the first six genetic principal components (PCs) as predictors. For binary classification (disease status), age was defined as the age at the end of follow-up. For age-of-onset prediction, age was defined as baseline age at recruitment. Second, best model - extended the null model by incorporating the single best-performing PGS, determined based on the highest ROC AUC in the training set. This selection process ensured that no data leakage occurred from the test set. Third, multi model - included the null model covariates along with an optimal combination of 1,514 PGSs, selected through elastic net regularization. This approach was applied to both binary classification and age-of-onset prediction. In most of the cases, the majority of PGSs after regularization had weight of zero, thus in the end only a small subset of PGSs were taken for model training.

For diagnosis prediction, we trained a logistic regression model with elastic net regularization (penalty=elasticnet), implemented in the scikit-learn (v.1.5.2) Python package. To

ensure proper model fitting and comparability across features, we applied standardization to all predictors before model training. Hyperparameter tuning was performed using a randomized search strategy, using ROC AUC as the scoring metric. We optimized: the inverse regularization strength (C) and L1 regularization ratio (l1_ratio). A total of 200 iterations with 3-fold cross-validation were conducted. The best-performing model from this search was then applied to the test set. Corresponding results were reported in the Supplementary Data 7.

For age-of-onset prediction, we trained a Cox proportional hazards model with elastic net regularization (CoxNet), implemented in the scikit-survival (v.0.23.0) Python package. As in the classification models, all features were standardized before training. Model selection and hyperparameter tuning followed the same randomized search approach, ensuring robust performance evaluation. The only difference was that we excluded patients with baseline age less than age of the disease onset or age of follow up, thus sample sizes between diagnosis prediction and age-of-onset differed and for each disease there was a unique number of patients satisfying this filtering procedure. We used years of follow up as a time scale.

Further, we systematically compared null, best, and multi models. We aimed to quantify the added predictive value, comparing, first, single best PGS with null model and, second, multiple PGSs with the null model. For binary prediction we utilized ROC AUC, for survival model we checked time-dependent ROC AUC, Brier Score and D-calibration measures (see Supplementary Materials, PGS evaluation measures). Corresponding results were reported in the Supplementary Data 8.

Feature importances, defined as an absolute value of the elastic net model coefficients, were extracted from Logistic Regression and CoxNet multi models to define optimal combination of PGSs. All scores with weight of zero were excluded. Corresponding weights were reported in the Supplementary Data 9,10.

Ancestry adjustment

Raw PGS values for individual patients can be biased by underlying population structure, so we implemented the adjustment method described in Hao et al.[11] to remove ancestry-driven variance. First, we regressed each raw PGS on genetic principal components (PCs) (see Supplementary Materials, Ancestry prediction in FinnGen cohort) to predict the ancestry-driven component (predicted PGS). For each of the 3,168 PGSs we used a 6-fold cross-validation approach to derive this component for every FinnGen participant. Next, we subtracted predicted PGS from the raw score. The resulting adjusted PGS retains the genetic signal while minimizing the variance attributed solely to population structure:

$$\mathrm{PGS}_{adjusted} = \mathrm{PGS}_{raw} - \mathrm{PGS}_{predicted} \quad (8)$$

Notably, all predicted PGSs exhibited a statistically significant correlation with their respective raw PGSs, confirming the influence of population structure on all of the PGSs (Figure 6c). Expectedly, predicted PGSs were completely explained by the PCs and adjusted PGSs were completely decorrelated with PCs (Supplementary Figure 10).

UK Biobank cohort

The UK Biobank (UKBB) dataset consists of 502,226 participants, Age at the time of recruitment ranges from 37 to 73, with median age of 58 years. 54% of them were females. Of these, we kept only 78,336 participants, who were classified as non-White British (NA at Data-Field 22006). All of these samples were projected onto 1000 Genomes principal components, previously used for FinnGen, and ancestry groups were inferred using previously used random forest classifier (see Supplementary Materials, Ancestry prediction in FinnGen cohorts; Supplementary Figure 17). As a result, 9,139 participants were classified as Africans (AFR), 4,639 as Admixed Americans (AMR), 2,447 as East Asians (EAS), 48,745 as Non-white British Europeans (EUR), 9,518 as South Asians (SAS) and 3,848 as Others. UKBB phenocodes were converted to FinnGen endpoints using the following script: https://github.com/FINNGEN/pan-ukbb-mapping.

Public predictive models

Initially, out of 110 tested best-PGS CoxNet models only 22 met our inclusion criteria for public release: individual survival curves had perfect calibration (D-calibration p-value > 0.99), time-dependent AUC (1-10 year) more than 0.65 on internal withhold FinnGen test set and increment in AUC from inclusion of PGS at least of 1% (Supplementary Figure 16). They were finetuned using the same procedures described previously (see Supplementary Materials, Disease risk prediction models). As an input they take biological sex, current age of the patient and adjusted PGS percentile, which can be calculated locally by the user. Thus, no PC information needed, or individual-level data transmitted to the browser.

Further, we validated these models trained on FinnGen on an external set of non-White British participants of UKBB (see Supplementary Materials, UK Biobank cohort). In such a way we both tested models on a diverse ancestral group and restricted testing on a cohort were not used for model development in Privé F. et al. (PGP000263).

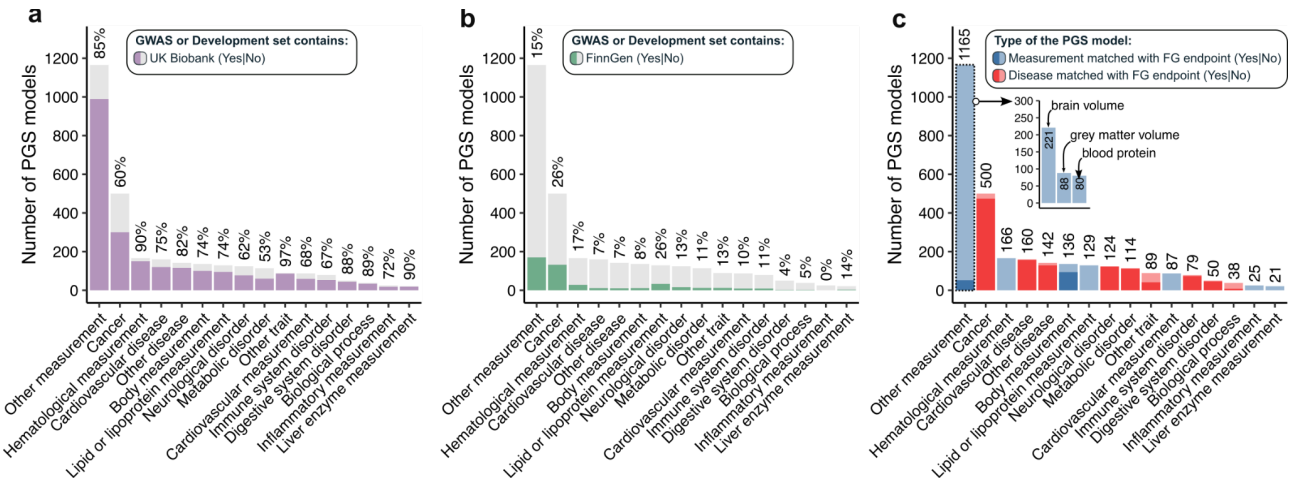
First, we assessed whether ancestry-adjustment models trained in FinnGen effectively removed population-structure effects in UK Biobank (Supplementary Fig. 18). UK Biobank participants were projected into the predefined 1000 Genomes principal component space, and the resulting ancestry PCs were used with the FinnGen-trained models to residualize the ancestry-driven component of each PGS, as described above for FinnGen. This procedure was implemented using pgsb-cli, which automates projection and adjustment.

21 out of 22 had variance explained by the first six principal components less than 0.05. Further, we derived ROC AUCs for binary separation of cases and controls by individual PGSs (Supplementary Figure 19) and time-dependent ROC AUC (1-15 years) for CoxNet models (Supplementary Figure 20), keeping those who had at least 3 major continental groups with AUC significantly higher than 0.6. As a result, only 11 models passed external validation on UKBB cohort (Supplementary Figure 21).

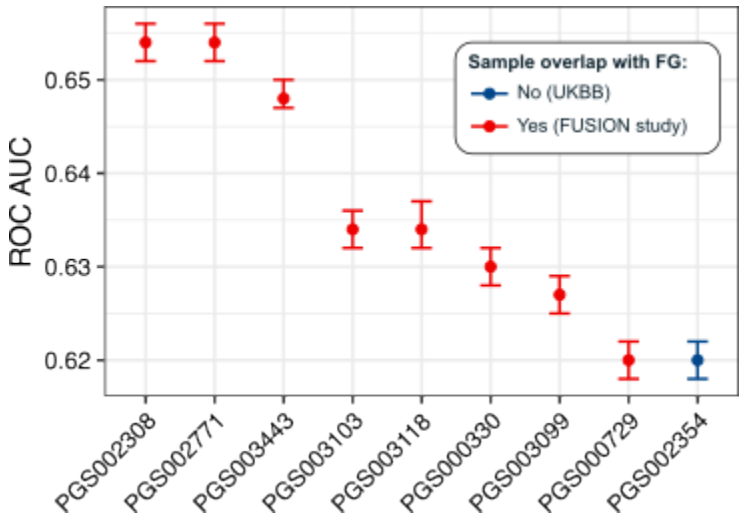
PGS Browser command-line tool

To automate percentile calculation for an ancestry-adjusted PGS in any individual, we developed pgsb-cli (see Documentation page at www.pgs.nchigm.org) - a Docker-based command-line tool that runs entirely on the user's machine: it takes as an input genotype data (vcf or plink format) and a PGS model downloaded from the PGS Browser, matches variants, projects the sample onto 1000 Genomes principal components, computes the predicted and adjusted PGSs, and

outputs standardized scores and percentiles, all without transmitting any individual-level data outside the local container.

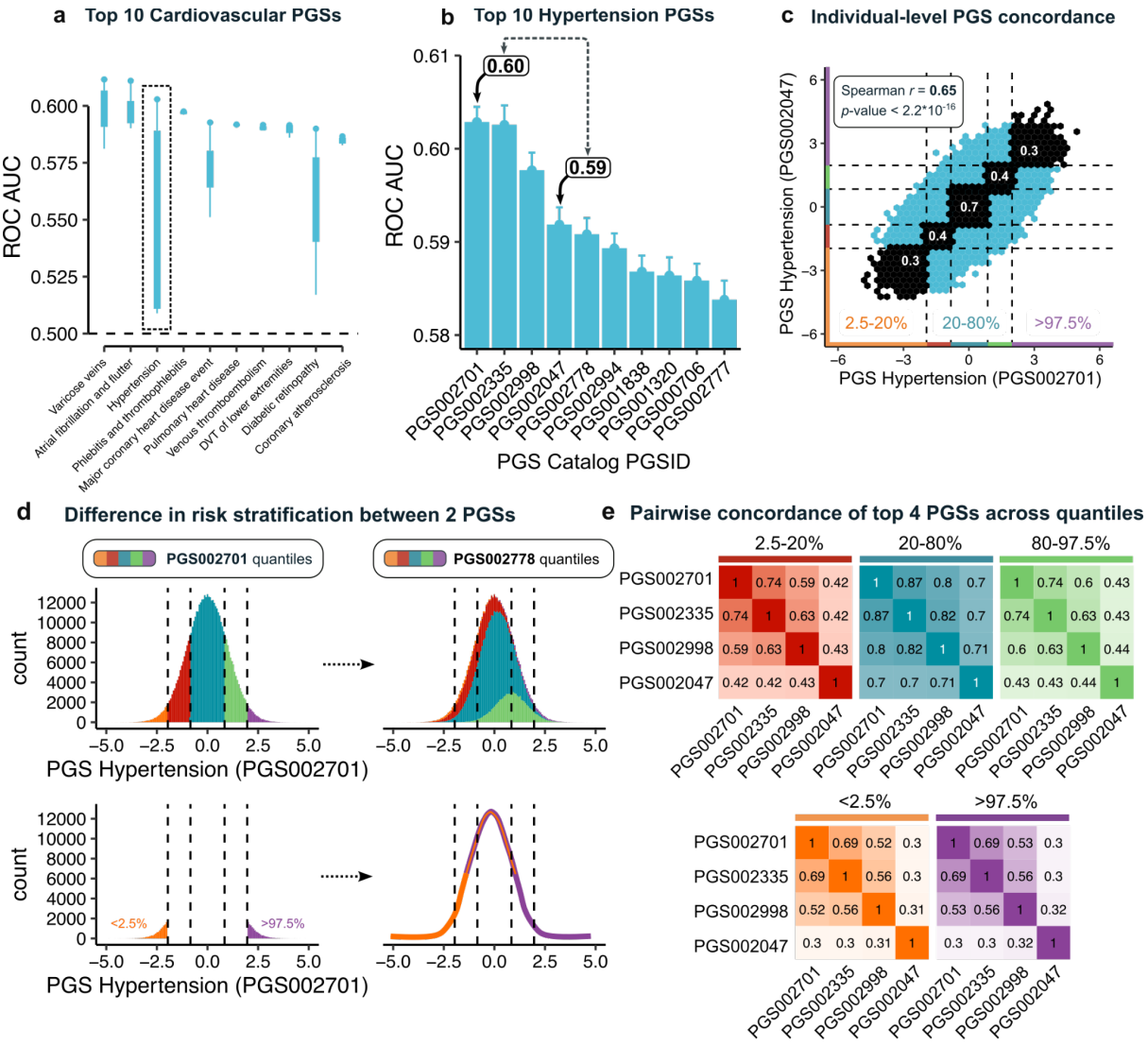


Supplementary Figure 1. PGS Models from the PGS Catalog included in the analysis. (a) Number of PGS models in the PGS Catalog that used the UK Biobank as a source of effect sizes or as a development set during model creation. Numbers above bars indicate percentage out of total number of scores in the corresponding category. (b) Number of PGS models in the PGS Catalog that used the FinnGen cohort as a source of effect sizes or as a development set during model creation. Numbers above bars indicate percentage out of total number of scores in the corresponding category. (c) Number of PGS models across trait categories, as classified in the PGS Catalog. Lighter bars indicate the total number of models in each category, whereas darker bars indicate the number of PGS models matched to a FinnGen endpoint. Numbers above the bars show the total number of scores in each category. Source data are provided as a Source Data file.

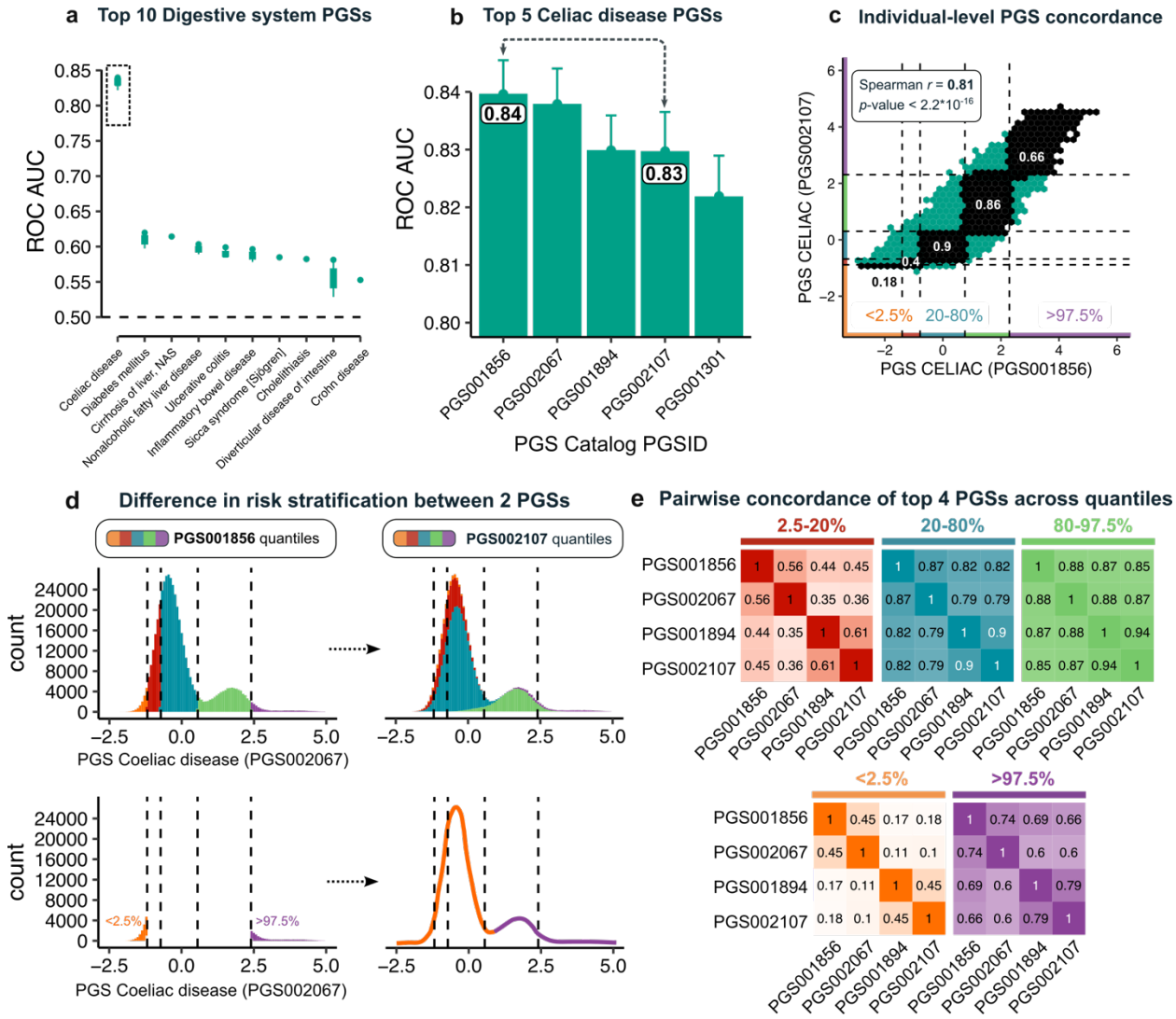


Supplementary Figure 2. Top 10 polygenic scores for type 2 diabetes. Each point represents the unadjusted ROC AUC estimated for one PGS, and horizontal error bars indicate 95% confidence intervals. ROC AUC and corresponding confidence intervals were estimated from 300 bootstrap resamples of the FinnGen dataset. Source data are provided as a Source Data file.

Supplementary Figure 3. Area under the receiver operating characteristic curve (ROC AUC) for 157 FinnGen disease endpoints. Dashed line depict ROC AUC equals 0.5. Color depicts broad disease category derived from PGS Catalog classification. Each point represents the unadjusted ROC AUC estimated for one PGS, and horizontal error bars indicate 95% confidence intervals. ROC AUC and corresponding confidence intervals were estimated from 300 bootstrap resamples of the FinnGen dataset. FinnGen disease endpoints were additionally preprocessed to depict true control distributions - all samples removed from controls in original FinnGen endpoint were reintroduced and treated as controls (see https://risteys.finngen.fi/). Total number of cases and controls for each endpoint are provided in Supplementary Data 2. M.n. - malignant neoplasm. Source data are provided as a Source Data file.

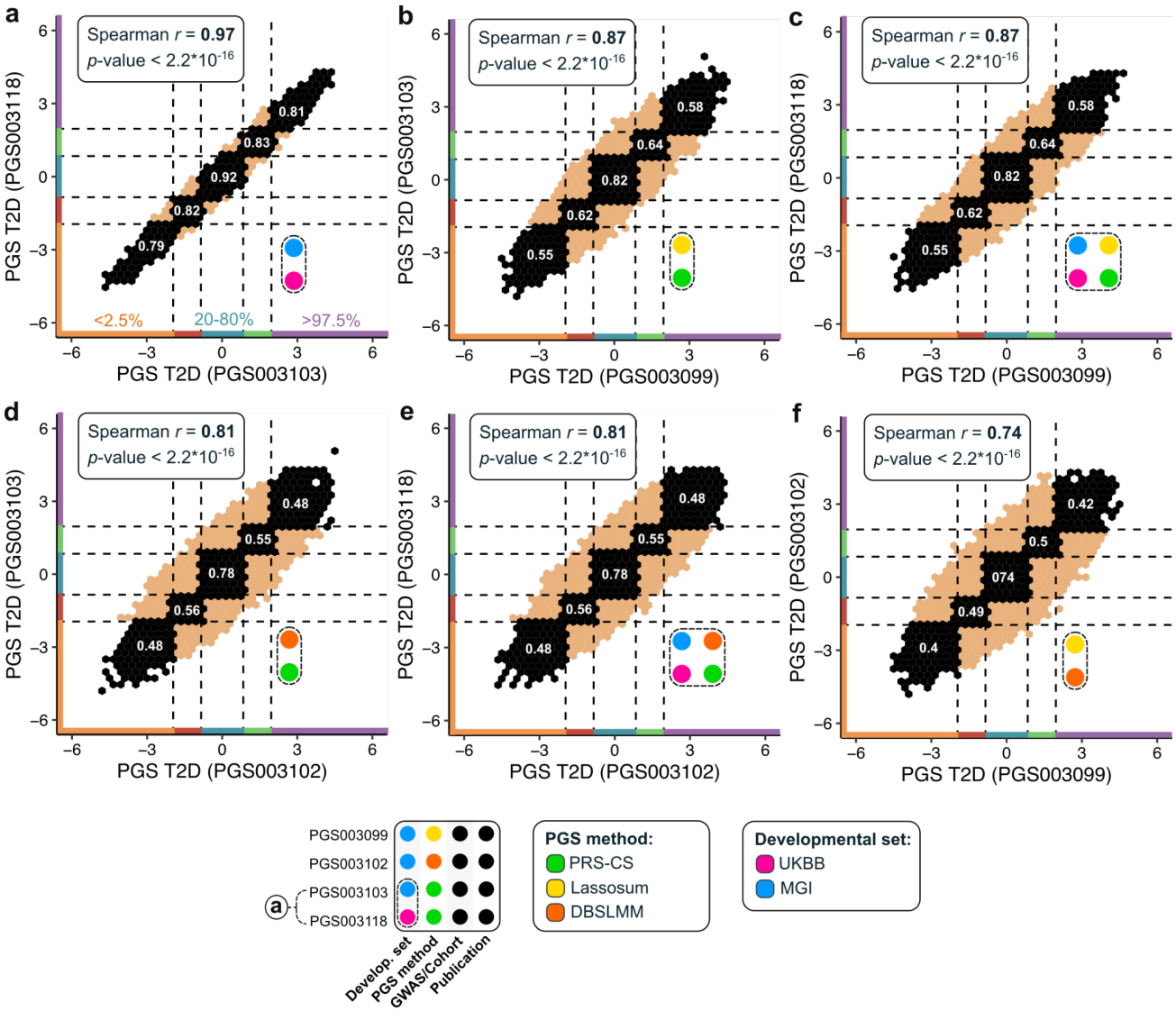


Supplementary Figure 4. Individual-level concordance between the top four Hypertension PGSs. (a) Top ten cardiovascular system PGSs. Boxplots summarize the distribution of ROC AUC values across distinct PGS models within each disease. Boxplots show interquartile range (box bounds), and 1.5 × interquartile range (whiskers); an individual dots indicate the model achieving the highest ROC AUC for the selected disease. The dashed box highlights the distribution of ROC AUCs for hypertension PGSs. Total number of PGS models for each disease can be found in Supplementary Data 1. (b) Top ten hypertension PGSs. The first and fourth PGSs are compared. Each point represents the unadjusted ROC AUC estimated for one PGS, and horizontal error bars indicate 95% confidence intervals. ROC AUC and corresponding confidence intervals were estimated from 300 bootstrap resamples of the FinnGen dataset. (c) Individual-level PGS concordance between the first and fourth PGSs from b. Colors indicate PGS distribution quantiles: orange (<2.5%), red (2.5-20%), turquoise (20-80%), green (80-97.5%), and violet (>97.5%). Dashed lines delineate the corresponding quantile regions. Numbers represent concordance between identical quantiles of the two PGSs. (d) Differences in risk stratification between the two PGSs, with colors depicting quantiles. (e) Pairwise concordance of the top four hypertension PGSs across quantiles.

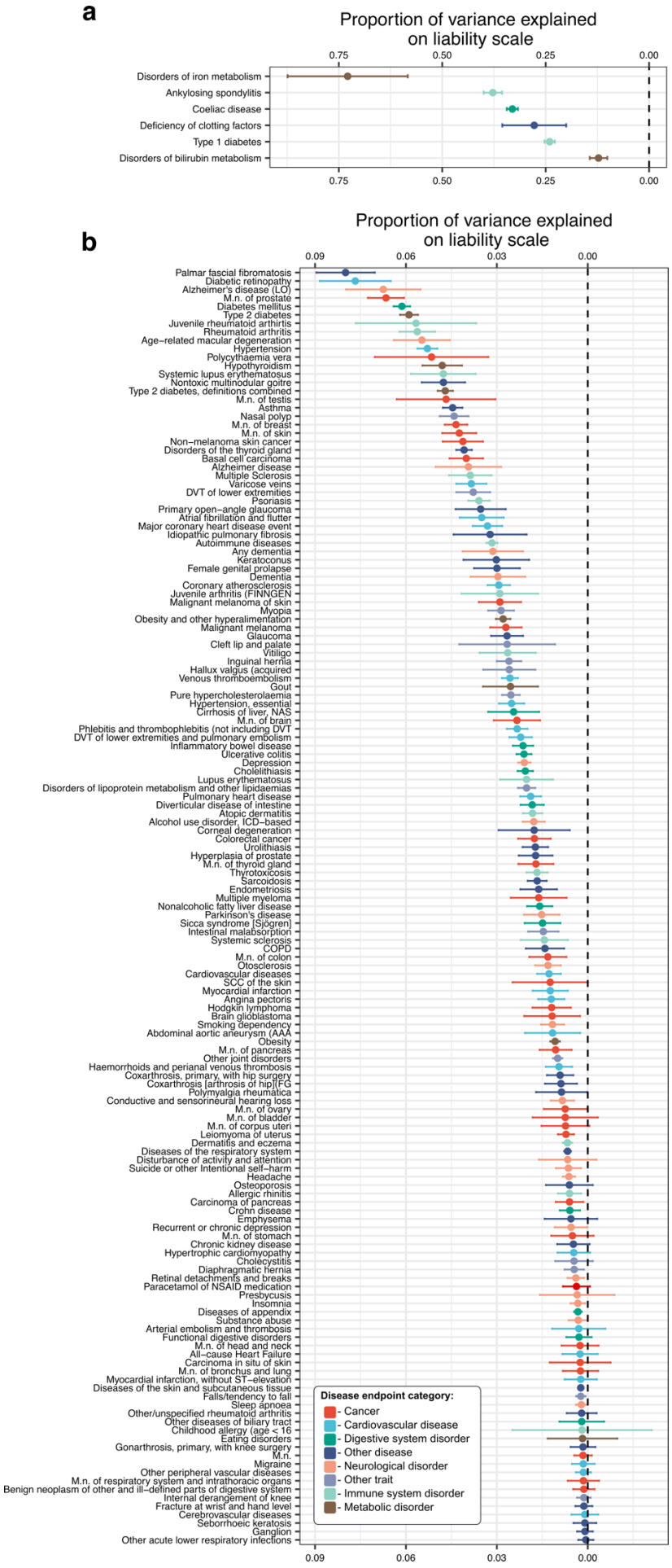


Supplementary Figure 5. Individual-level concordance between the top four Celiac disease PGSs. (a) Top ten digestive system PGSs. Boxplots summarize the distribution of ROC AUC values across distinct PGS models within each disease. Boxplots show interquartile range (box bounds), and 1.5 × interquartile range (whiskers); an individual dots indicate the model achieving the highest ROC AUC for the selected disease. The dashed box highlights the distribution of ROC AUCs for Celiac disease PGSs. Total number of PGS models for each disease can be found in Supplementary Data 1. (b) Top five Celiac disease PGSs, no statistically significant differences among the top four were found (two-sided Walds z-test p-value > 0.05). The first and fourth PGSs are compared. Each point represents the unadjusted ROC AUC estimated for one PGS, and horizontal error bars indicate 95% confidence intervals. ROC AUC and corresponding confidence intervals were estimated from 300 bootstrap resamples of the FinnGen dataset. (c) Individual-

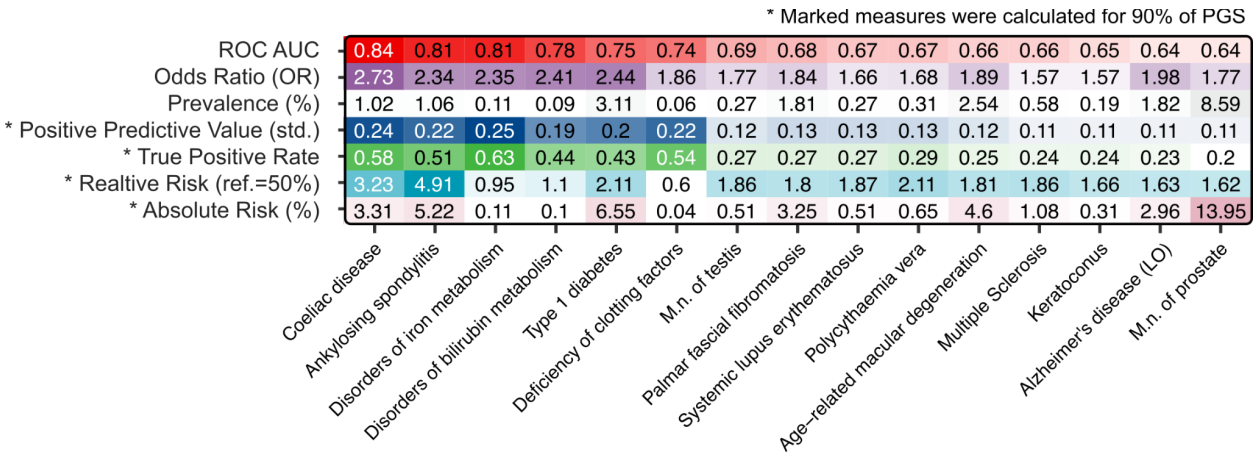
level PGS concordance between the first and fourth PGSs from b. Colors indicate PGS distribution quantiles: orange (<2.5%), red (2.5-20%), turquoise (20-80%), green (80-97.5%), and violet (>97.5%). Dashed lines delineate the corresponding quantile regions. Numbers represent concordance between identical quantiles of the two PGSs. (d) Differences in risk stratification between the two PGSs, with colors depicting quantiles. (e) Pairwise concordance of the top four Celiac disease PGSs across quantiles.



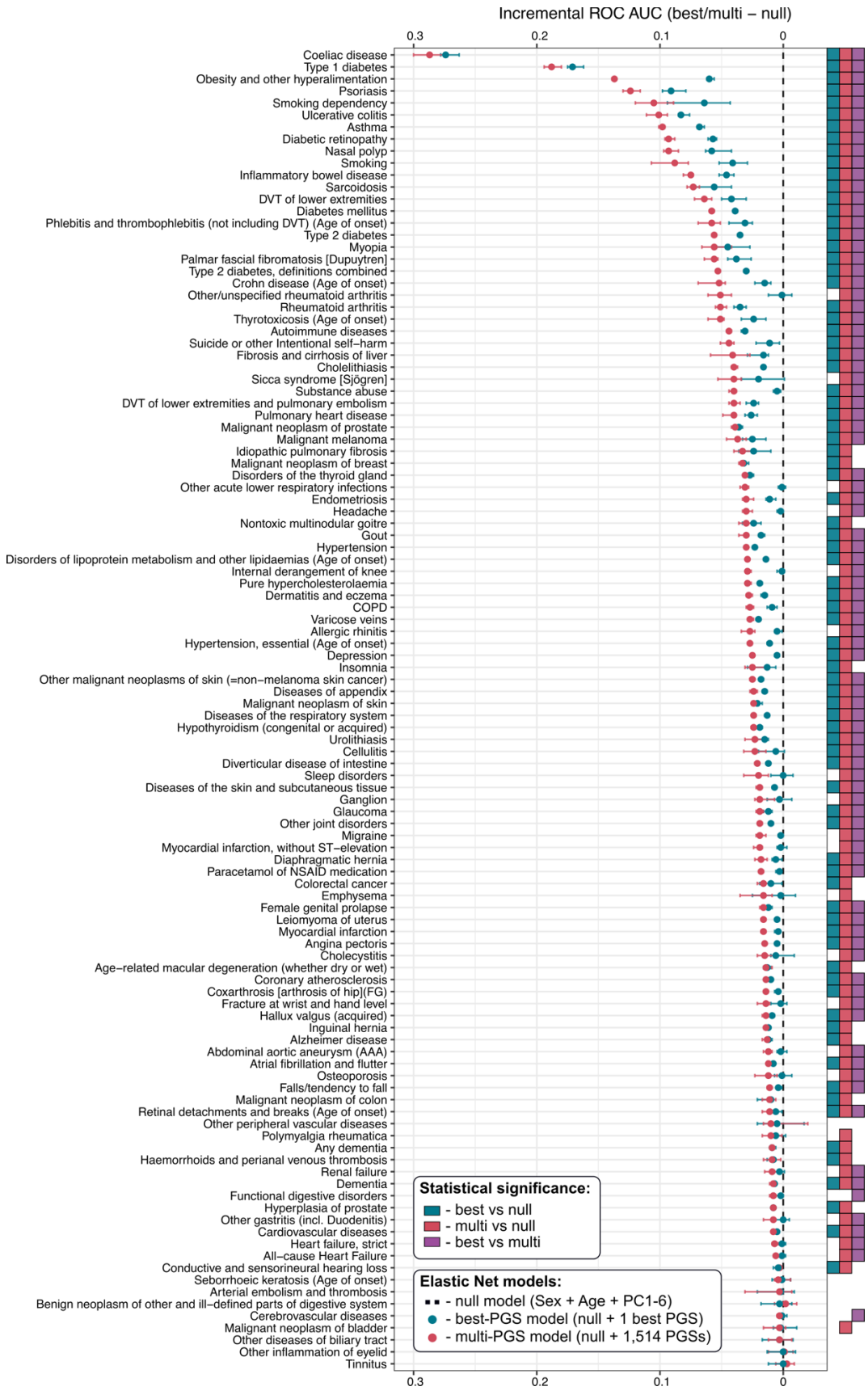
Supplementary Figure 6. Individual-level concordance among the top four type 2 diabetes PGSs derived from the same GWAS. The legend indicates the methodological combinations used to construct each score. Although all four scores were derived from the same GWAS data, methodological differences produced substantial discordance in risk stratification. Numbers within overlapping quantile segments indicate the proportion of individuals shared between the corresponding quantiles. Dashed lines delineate the quantile regions: <2.5%, 2.5-20%, 20-80%, 80-97.5% and >97.5%.



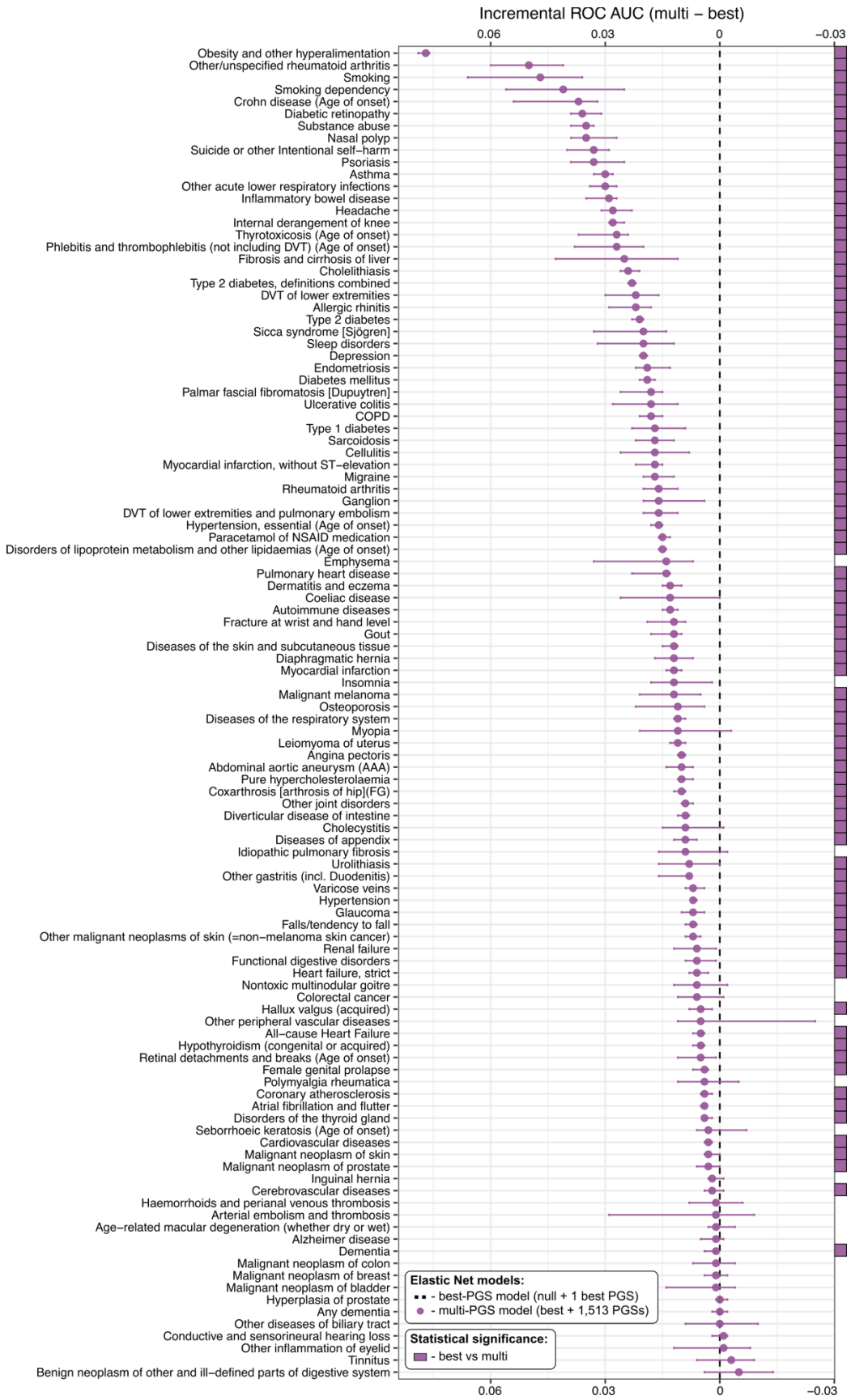
Supplementary Figure 7. Proportion of variance explained by PGS on liability scale for 157 FinnGen disease endpoints. Points represent liability-scale R2 values and error bars indicate 95% confidence intervals estimated from 50 bootstrap resamples of the endpoint-specific FinnGen data sets. Total number of cases and controls for each endpoint are provided in Supplementary Data 4. (a) Top 6 endpoints. (b) Remaining 151 endpoints. Dashed line depicts variance equals 0. Color depicts broad disease category derived from PGS Catalog classification. FinnGen disease endpoints were additionally preprocessed to depict true control distributions - all samples removed from controls in original FinnGen endpoint were reintroduced and treated as controls. M.n. - malignant neoplasm. Source data are provided as a Source Data file.



Supplementary Figure 8. Various predictive performance metrics for top 15 scores. PGSs are sorted based on ROC AUC. M.n., malignant neoplasm. Source data are provided as a Source Data file.



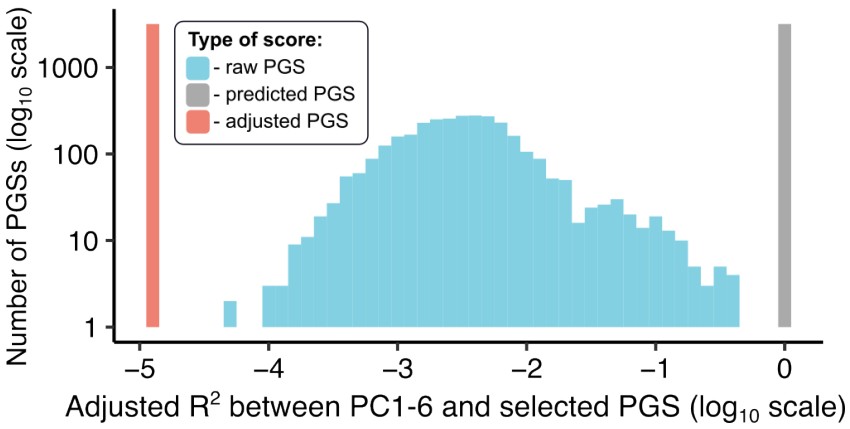
Supplementary Figure 9. Incremental area under the receiver operating characteristic curve (ROC AUC) for three elastic-net models. The null model included sex, age at the end of follow-up, and the first six principal components as predictors. Incremental ROC AUC values were defined relative to the null model and are therefore equal to zero for the null model across all phenotypes (dashed line). The best-PGS model included the predictors from the null model together with the single best-performing PGS for the selected phenotype. The multi-PGS model included the predictors from the null model together with 1,514 PGSs. Each point represents ROC AUC for one binary FinnGen endpoint, and error bars indicate 95% confidence intervals estimated from 150 bootstrap resamples of the endpoint-specific FinnGen test sets. Colored squares indicate statistical significance for three pairwise comparisons: best-PGS versus null; multi-PGS versus best-PGS. Significance of differences in ROC AUC was assessed using a one-sided paired bootstrap test. The presence of a square to the right of a given phenotype indicates a significant ROC AUC difference in the test set for the corresponding comparison. Exact case and control counts for the train and test sets for each endpoint are provided in the corresponding Source Data file. M.n., malignant neoplasm. Source data are provided as a Source Data file.



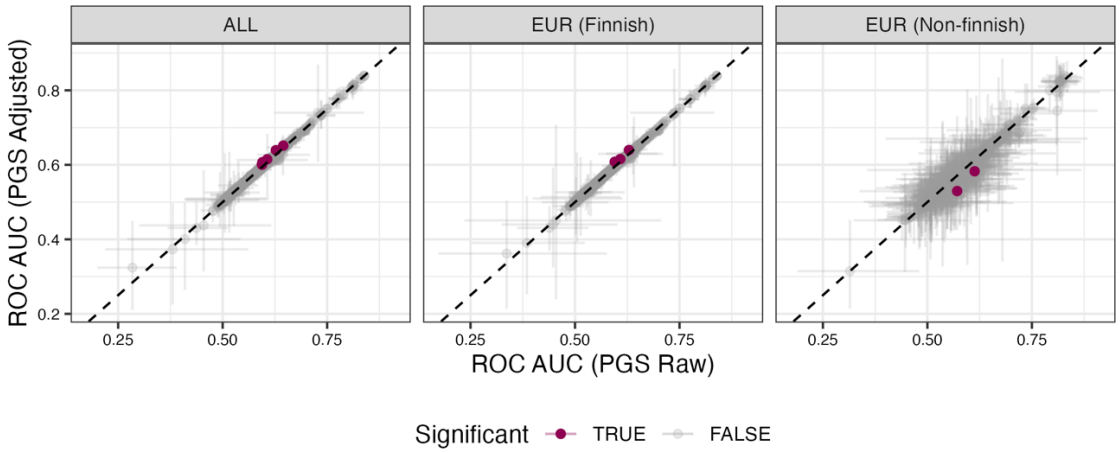
Supplementary Figure 10. Incremental area under the receiver operating characteristic curve (ROC AUC) for two elastic-net models. The best PGS model included sex, age at the end of follow-up, the first six principal components and single PGS best for target phenotype as predictors. Incremental ROC AUC values were defined relative to the best PGS model and are therefore equal to zero for the best model across all phenotypes (dashed line). The multi-PGS model included the predictors from the null model (only sex, age at baseline and first six PCs) together with 1,514 PGSs. Each point represents ROC AUC for one binary FinnGen endpoint, and error bars indicate 95% confidence intervals estimated from 150 bootstrap resamples of the endpoint-specific FinnGen test sets. Colored squares indicate statistical significance for pairwise comparisons between multi-PGS versus best-PGS ROC AUCs. Significance of differences in ROC AUC was assessed using a one-sided paired bootstrap test. The presence of a square to the right of a given phenotype indicates a significant ROC AUC difference in the test set for the corresponding comparison. Exact case and control counts for the train and test sets for each endpoint are provided in the corresponding Source Data file. M.n., malignant neoplasm. Source data are provided as a Source Data file.

Supplementary Figure 11. Incremental time-dependent (td) (1-10 years) area under the receiver operating characteristic curve (ROC AUC) between three Cox's proportional hazard's models with elastic net penalty (Cox Net). The null model included sex, age at the end of follow-up, and the first six principal components as predictors. Incremental ROC AUC values were defined relative to the null model and are therefore equal to zero for the null model across all phenotypes (dashed line). The best-PGS model included the predictors from the null model together with the single best-performing PGS for the selected phenotype. The multi-PGS model included the predictors from the null model together with 1,514 PGSs. Each point represents ROC AUC for one binary FinnGen endpoint, and error bars indicate 95% confidence intervals estimated from 150 bootstrap resamples of the endpoint-specific FinnGen test sets. Colored squares indicate statistical significance for three pairwise comparisons: best-PGS versus null; multi-PGS versus null; multi-PGS versus best-PGS. Significance of differences in ROC AUC was assessed using a one-sided paired bootstrap test. The presence of a square to the right of a given phenotype indicates a significant ROC AUC difference in the test set for the corresponding comparison. Exact case and control counts for the train and test sets for each endpoint are provided in the corresponding Source Data file. M.n., malignant neoplasm. Source data are provided as a Source Data file.

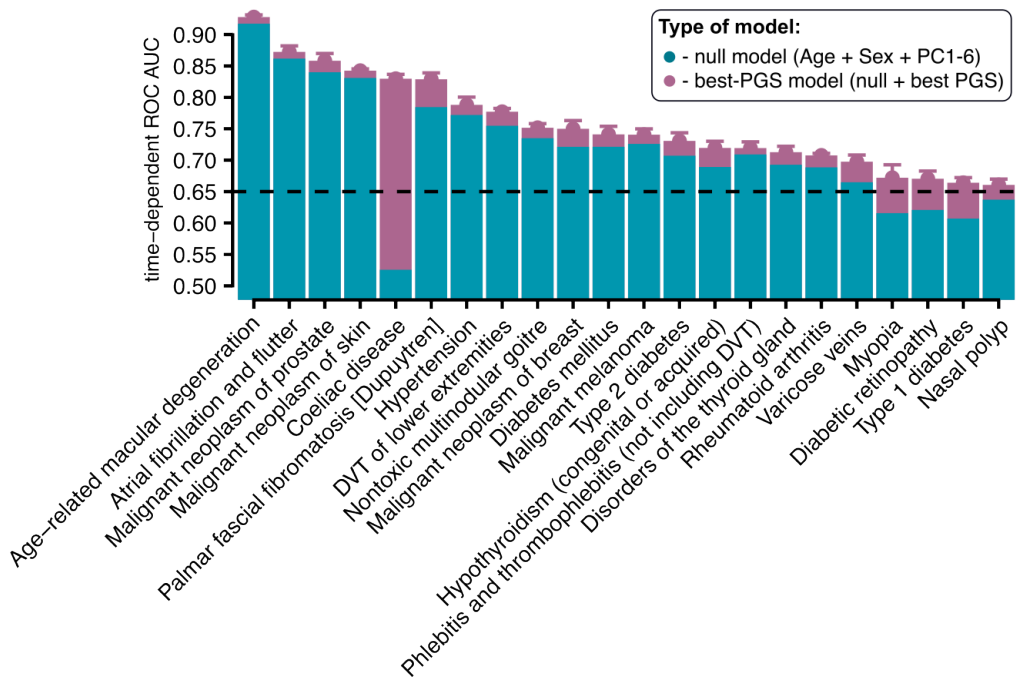
Supplementary Figure 12. Incremental time-dependent (td) (1-10 years) area under the receiver operating characteristic curve (ROC AUC) for two Cox's proportional hazard's models with elastic net penalty (Cox Net). The best PGS model included sex, age at baseline, the first six principal components and single PGS best for target phenotype as predictors. Incremental ROC AUC values were defined relative to the best PGS model and are therefore equal to zero for the best model across all phenotypes (dashed line). The multi-PGS model included the predictors from the null model (only sex, age at baseline and first six PCs) together with 1,514 PGSs. Each point represents ROC AUC for one binary FinnGen endpoint, and error bars indicate 95% confidence intervals estimated from 150 bootstrap resamples of the endpoint-specific FinnGen test sets. Colored squares indicate statistical significance for pairwise comparisons between multi-PGS versus best-PGS ROC AUCs. Significance of differences in ROC AUC was assessed using a one-sided paired bootstrap test. The presence of a square to the right of a given phenotype indicates a significant ROC AUC difference in the test set for the corresponding comparison. Exact case and control counts for the train and test sets for each endpoint are provided in the corresponding Source Data file. M.n., malignant neoplasm. Source data are provided as a Source Data file.



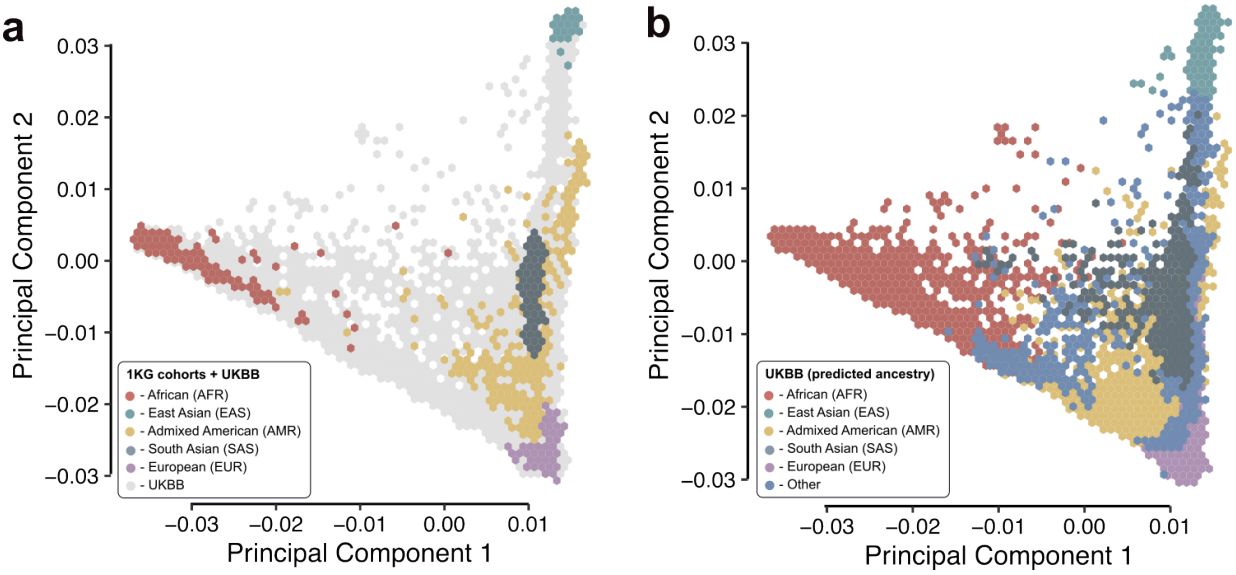
Supplementary Figure 13. Lack of correlation between adjusted PGS and PCs. Both X and Y-axis are in $\log_{10}$ scale. Source data are provided as a Source Data file.



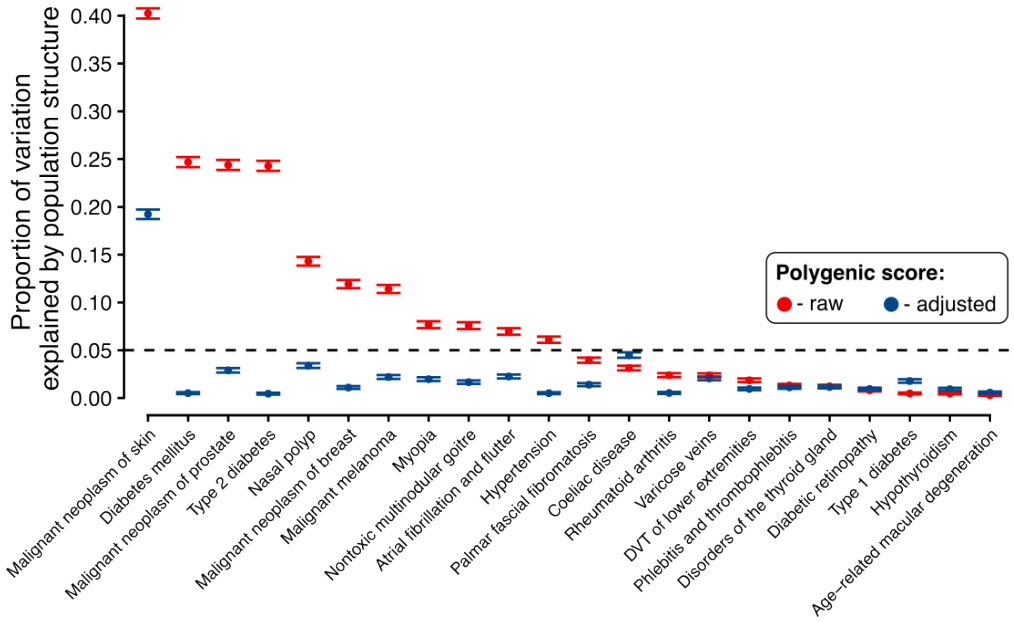
Supplementary Figure 14. Comparison of ROC AUC values for raw and ancestry-adjusted PGSs across European cohorts (Finnish-only, non-Finnish and combined dataset). Each point represents the unadjusted ROC AUC estimated for one PGS, and horizontal error bars indicate 95% confidence intervals. ROC AUC and corresponding confidence intervals were estimated from 300 bootstrap resamples of the FinnGen dataset. Points above the dashed diagonal line indicate higher accuracy after adjustment, whereas points below the line indicate reduced accuracy. Colored points highlight scores significantly deviating from the diagonal, indicating a difference in ROC AUC. To test difference in RO AUCs for significance we used two-sided Wald-type z-test, followed by Benjamini-Hochberg correction for multiple comparison. Groups: Finnish Europeans (EUR Finnish), non-Finnish Europeans (EUR non-Finnish), combined cohort (ALL). Source data are provided as a Source Data file.



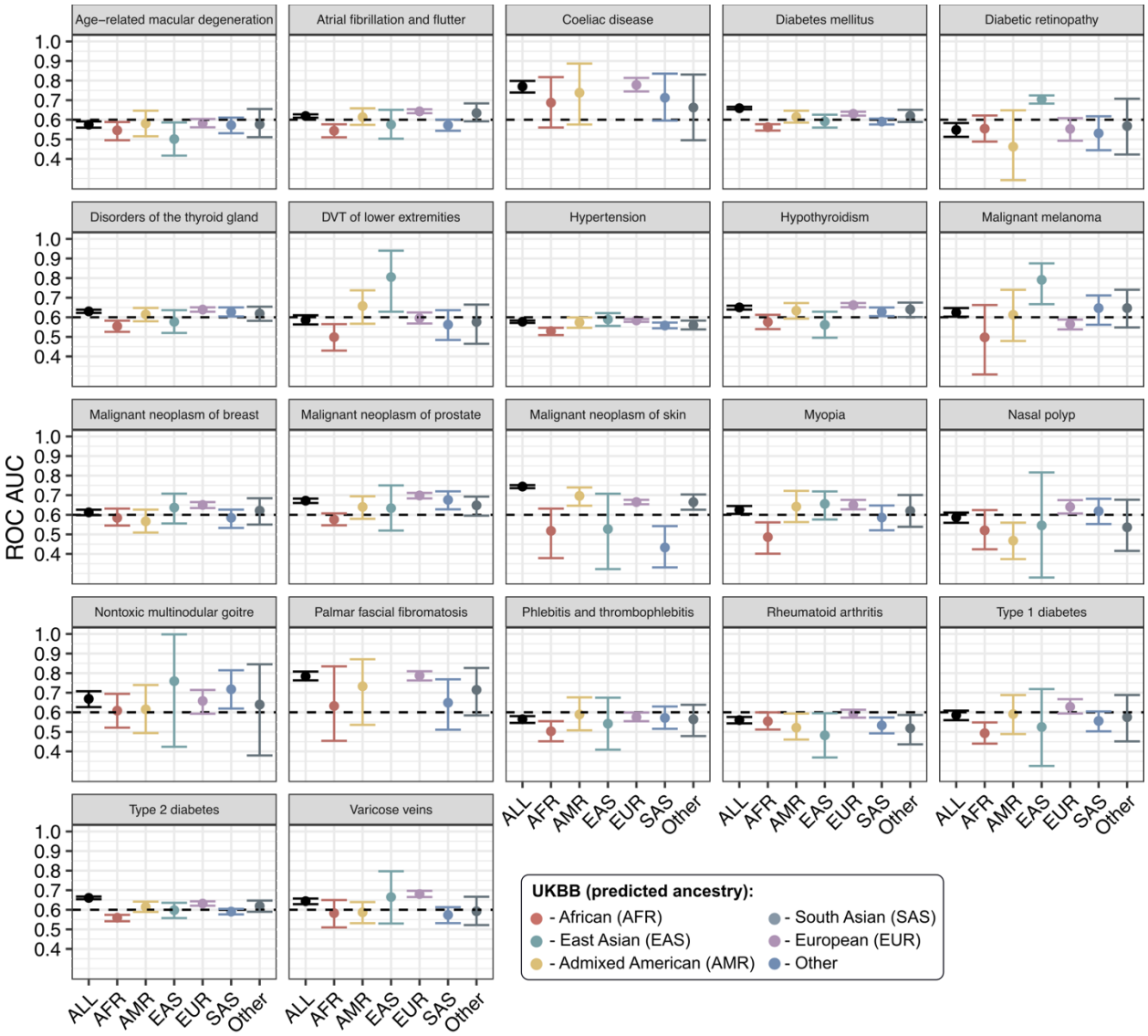
Supplementary Figure 15. Time-dependent ROC AUC values for the 22 externally released CoxNet models evaluated in the FinnGen test set. Each point represents the mean time-dependent ROC AUC across years 1-10 for the corresponding best single-PGS CoxNet model, and error bars indicate 95% confidence intervals estimated from 150 bootstrap resamples of the endpoint-specific FinnGen test sets. Exact case and control counts for each endpoint are reported in Supplementary Data 8. The dashed line depicts the ROC AUC threshold (0.65) used for initial selection for external validation. Source data are provided as a Source Data file.



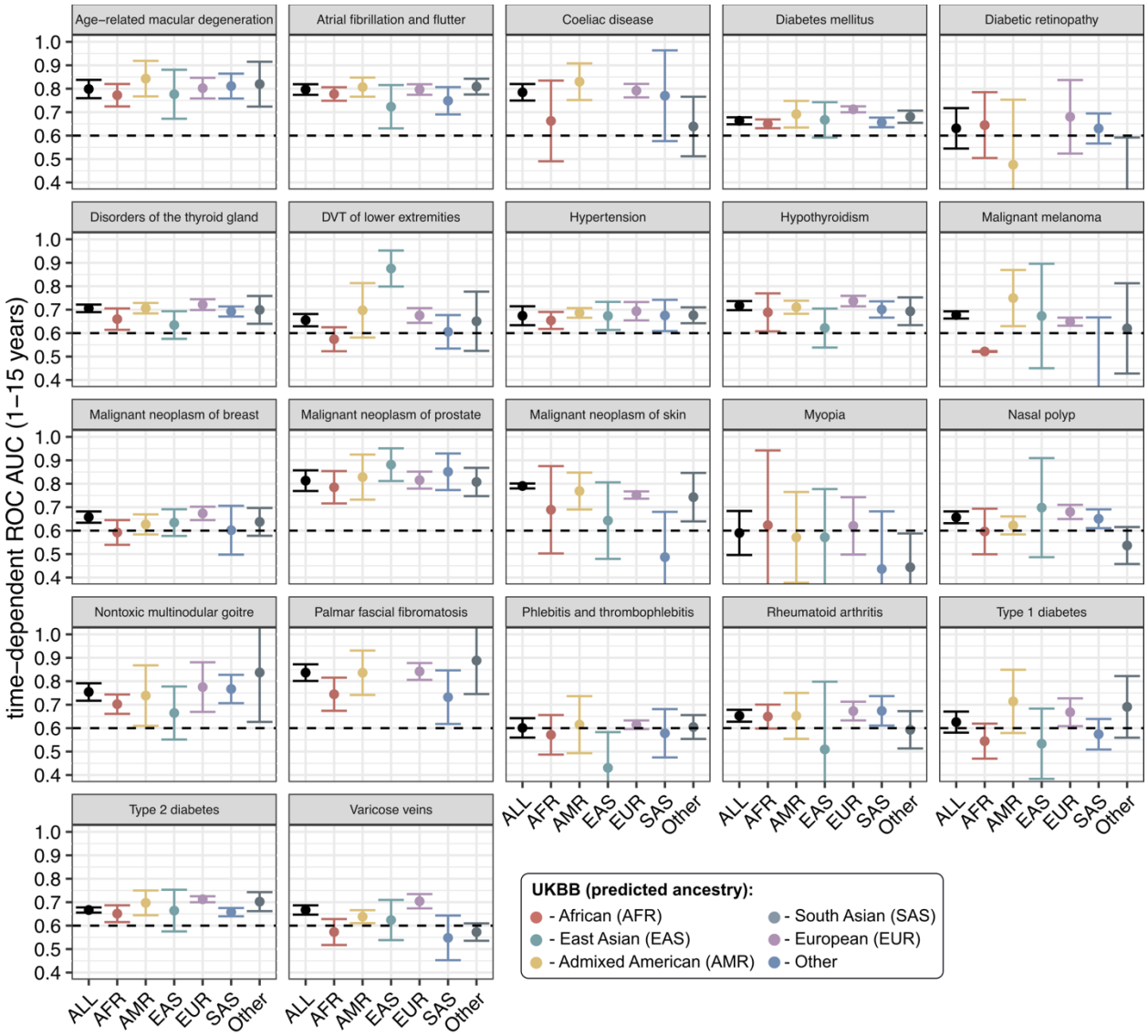
Supplementary Figure 16. (a) UK biobank (UKBB) projected onto 1000G (1KG) principal component coordinates. (b) Predicted ancestries for the UKBB cohort.



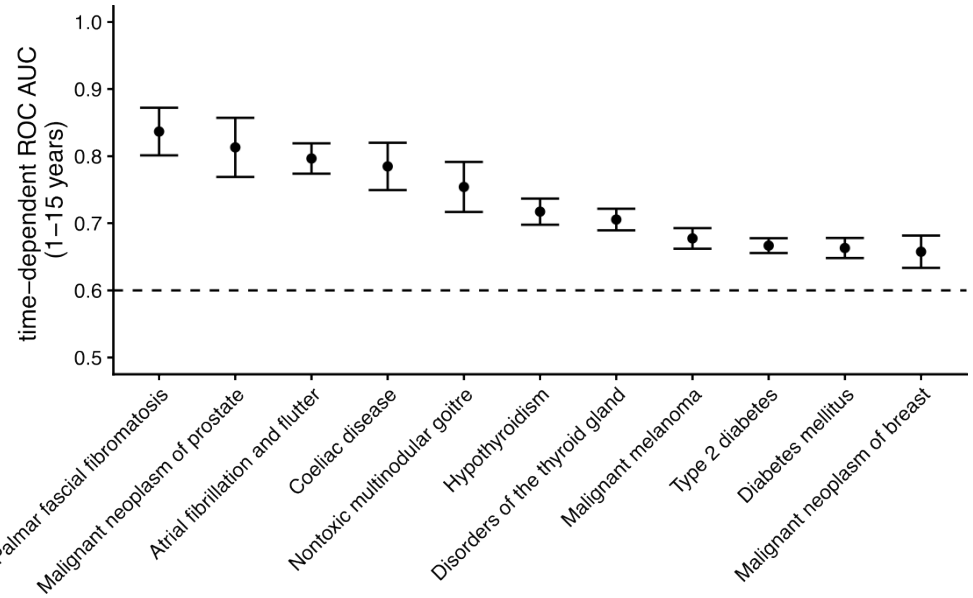
Supplementary Figure 17. Proportion of PGS variance explained by population structure for raw and ancestry-adjusted scores in the UK Biobank (UKBB) test cohort. Adjustment parameters were estimated in FinnGen. Points represent R^2 values, and error bars indicate 95% confidence intervals derived from linear regression models. Source data are provided as a Source Data file.



Supplementary Figure 18. ROC AUCs for the 22 PGSs incorporated into selected single-PGS CoxNet models, evaluated on the UK Biobank (UKBB) dataset. Points represent mean time-dependent ROC AUC values, and error bars indicate 95% confidence intervals estimated from 300 bootstrap resamples. The dashed line denotes the ROC AUC threshold of 0.6. To assess whether ROC AUC values differed significantly from 0.6, we used a one-sided Wald-type z-test. Exact case and control counts for test sets for each endpoint are provided in the corresponding source file. Source data are provided as a Source Data file.



Supplementary Figure 19. Time-dependent ROC AUC values (1-15 years) for 22 CoxNet models trained in FinnGen and externally validated in the UK Biobank (UKBB) cohort. Points represent mean time-dependent ROC AUC values, and error bars indicate 95% confidence intervals estimated from 150 bootstrap resamples. The dashed line denotes the ROC AUC threshold of 0.6. To assess whether ROC AUC values differed significantly from 0.6, we used a one-sided Wald-type z-test. Exact case and control counts for test sets for each endpoint are provided in in the corresponding source file. Source data are provided as a Source Data file.



Supplementary Figure 20. Time-dependent ROC AUC values across the 1-15 year prediction horizon for 11 CoxNet models trained in FinnGen, externally validated in UK Biobank (UKBB), and released through the PGS Browser. Points represent mean time-dependent ROC AUC values, and error bars indicate 95% confidence intervals estimated from 150 bootstrap resamples. The dashed line marks the ROC AUC threshold of 0.6. Source data are provided as a Source Data file.

References

1. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).
2. Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A better coefficient of determination for genetic profile analysis. *Genet Epidemiol* **36**, 214–224 (2012).
3. Heston, T. F. Standardizing predictive values in diagnostic imaging research. *J Magn Reson Imaging* **33**, 505; author reply 506–7 (2011).
4. Haider, H., Hoehn, B., Davis, S. & Greiner, R. Effective ways to build and evaluate individual survival distributions. *J. Mach. Learn. Res.* **21**, (2020).
5. Tanigawa, Y. *et al.* Significant sparse polygenic risk scores across 813 traits in UK Biobank. *PLoS Genet* **18**, e1010105 (2022).
6. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am J Hum Genet* **109**, 373 (2022).
7. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat Genet* **54**, 450–458 (2022).
8. Liu, N. *et al.* Cross-ancestry genome-wide association meta-analyses of hippocampal and subfield volumes. *Nat Genet* **55**, 1126–1137 (2023).
9. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).
10. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet* **53**, 185–194 (2021).
11. Hao, L. *et al.* Development of a clinical polygenic risk score assay and reporting workflow. *Nat Med* **28**, 1006–1013 (2022).
12. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).