Supplementary Information

Table of Contents

Supplementary Notes

Supplementary Note 1. PGS Catalog Scoring File Specifications.

The PGS Catalog's Scoring File format is described on our website: https://www.pgscatalog.org/downloads/. Each scoring file (variant information, effect alleles/weights) is formatted to be a gzipped tab-delimited text file, labelled by its PGS Catalog Score ID (e.g. `PGS000001.txt.gz`). We developed the scoring file format to closely resemble existing formats used to calculate scores in common software (e.g. PLINK) so that users could easily apply these scores within existing pipelines.

Scores are extracted from the relevant publication, and a consistent header (lines starting with `#`) has been added to each file listing relevant information about the PGS with links to the original publication and Catalog identifier:

```
### PGS CATALOG SCORING FILE - see www.pgscatalog.org/downloads/#dl_ftp for
additional information
## POLYGENIC SCORE (PGS) INFORMATION
# PGS ID = PGS identifier, e.g. 'PGS000001'
# PGS Name = PGS name, e.g. 'PRS77_BC' - optional
# Reported Trait = trait, e.g. 'Breast Cancer'
# Original Genome Build = Genome build/assembly, e.g. 'GRCh38'
# Number of Variants = Number of variants listed in the PGS
## SOURCE INFORMATION
# PGP ID = PGS publication identifier, e.g. 'PGP000001'
# Citation = Information about the publication
# LICENSE = License and terms of PGS use/distribution - refers to the EMBL-EBI
Terms of Use by default
rsID chr_name chr_position effect_allele reference_allele...
```

PGS scoring files are re-formatted to have consistent column headings based on the following schema:

Column Header	Field Name	Field Description	Mandatory?
rsID	dbSNP Accession ID (rsID)	The SNP's rs ID	**Yes -** Each PGS Scoring file must have either an *rsID* column or both a *chr_name* and *chr_position* column to identify the variant.
chr_name	Location - Chromosome	Chromosome name/number associated with the variant	
chr_position	Location - Base pair position within the Chromosome	Chromosomal position associated with the variant	
effect_allele	Effect Allele	The allele that's dosage is counted (e.g. {0, 1, 2}) and multiplied by the variant's weight ('effect_weight') when calculating score. The effect allele is also known as the 'risk allele'.	**Yes**
reference_allele	Reference Allele	The other allele(s) at the loci	**Suggested** - most software requires this for the calculation of scores and matching of the variants to existing genotype data,
effect_weight	Variant Weight	Value of the effect that is multiplied by the dosage of the effect allele ('effect_allele') when calculating the score.	**Yes**
locus_name	Locus Name	This is kept in for loci where the variant may be referenced by the gene (APOE e4). It is also common (usually in smaller PGS) to see the variants named according to the genes they impact.	*Optional*
weight_type	Type of Weight	Whether the author supplied Variant Weight is a: beta (effect size), or a log(OR/HR (odds/hazard ratio))	*Optional*

allelefrequency_effect	Effect Allele Frequency	Reported effect allele frequency, if the associated locus is a haplotype then haplotype frequency will be extracted.	*Optional*
is_interaction	FLAG: Interaction	This is a TRUE/FALSE variable that flags whether the weight should be multiplied with the dosage of more than one variant. Interactions are demarcated with a _x_ between entries for each of the variants present in the interaction.	*Optional*
is_recessive	FLAG: Recessive Inheritance Model	This is a TRUE/FALSE variable that flags whether the weight should be added to the PGS sum only if there are 2 copies of the effect allele (e.g. it is a recessive allele).	*Optional*
is_haplotype	FLAG: Haplotype or Diplotype	This is a TRUE/FALSE variable that flags whether the effect allele is a haplotype/diplotype rather than a single SNP. Constituent SNPs in the haplotype are semi-colon separated.	*Optional*
is_diplotype			
imputation_method	Imputation Method	This describes whether the variant was specifically called with a specific imputation or variant calling method. This is mostly kept to describe HLA-genotyping methods (e.g. flag SNP2HLA, HLA*IMP) that gives alleles that are not referenced by genomic position.	*Optional*
variant_description	Variant Description	This field describes any extra information about the variant (e.g. how it is genotyped or scored) that cannot be captured by the other fields.	*Optional*
inclusion_criteria	Score Inclusion Criteria	Explanation of when this variant is included into the PGS (e.g. if it depends on the results from other variants).	Optional

Supplementary Note 2. Inclusion Criteria for the PGS Catalog.

For the current PGS Catalog inclusion criteria see: https://www.pgscatalog.org/about. For a publication's data to be included in the PGS Catalog, it must fulfil the following criteria for either a newly developed polygenic score or an evaluation of an existing score(s):

A newly developed PGS

This includes the following information about the score and its predictive ability (evaluated on samples not used in training):

- Variant information necessary to apply the PGS to new samples (variant rsID and/or genomic position, weights/effect sizes, effect allele, genome build).
- Information about how the PGS was developed (computational method, variant selection, relevant parameters).

- Descriptions of the samples used for training (e.g. discovery of the variant associations [these can usually be extracted directly from the GWAS Catalog using GCST IDs], as well as fitting the PGS) and external evaluation.
- Establishment of the PGS' analytic validity, and a description of its predictive performance (e.g. effect sizes [beta, OR, HR, etc.], classification accuracy, proportion of the variance explained (R^2), and/or covariates evaluated in the PGS prediction).

An evaluation of a previously developed PGS

This would include the evaluation of PGS already present in the Catalog (or one that meets the inclusion criteria specified above), on samples not used for PGS training. The requirements for description would be the same as for the evaluation of a new PGS.

Supplementary Note 3. PGS Catalog Data Acquisition and Curation Processes.

The current PGS Catalog employs a manual search process to identify publications that may be eligible for inclusion in the PGS Catalog. Papers are identified on Google Scholar, PubMed and Twitter using common keywords: “genetic risk”, “polygenic risk”, “polygenic risk score”, “polygenic score”, and “genetic risk score”. A curator then scans the abstract and methods/results to identify whether the paper develops and/or validates a PGS (inclusion criteria #1), and adds it to our curation queue if the paper appears eligible. Subsequently, the paper is checked for the inclusion of PGS information (e.g. variants, effect alleles/weights), often sourced from supplementary excel spreadsheets within the paper, or in many cases extracted from external websites, figshare accessions, or Google drives linked within the paper. PGS information was determined to be available for inclusion in the Catalog provided no terms or restrictions on the data are imposed for download or resharing. If the data is unavailable, or sufficient information is not provided, the paper is marked as currently ineligible and a data-request email is sent to the corresponding authors if the paper is prioritized for curation. Papers suggested by users or communicated to us by authors are also checked according to this process and added to the queue.

Papers for full inclusion in the Catalog are selected from the list of eligible papers, prioritizing the papers that have been submitted to us by authors/users and based on data availability, citations, and our efforts to make the Catalog more comprehensive with respect to the diversity of traits included as well as ancestral diversity of populations represented in score development and evaluation. Full curation involves filling out a curation template (current version: www.pgscatalog.org/template/current) and formatting the variant information to have column headings consistent with our PGS Scoring File specification. Guidelines for filling out the curation template and extracting relevant data are provided online (current version: http://www.pgscatalog.org/docs/curation), and were developed in collaboration with experienced curators from the NHGRI-EBI GWAS Catalog. The curation guidelines describe the aspects of PGS study design captured in the Catalog, and each of the extracted data fields at each stage. Curation templates and scoring files were completed by expert curators, according to the information provided in the publication, or submitted by authors. All completed templates were validated by a second curator to ensure consistency before being

uploaded to the database. The most up to date description of the process for users to submit PGS data is provided at www.pgscatalog.org/submit.

The PGS Catalog data is released as available, but will move to a more regular release schedule in the future as data input increases. Individual PGS metadata or scores are versioned by date and provided as *archived_versions* on our FTP site (http://ftp.ebi.ac.uk/pub/databases/spot/pgs/) along with *previous_releases* of the complete Catalog metadata if any changes are made.

Supplementary Note 4. PGS Catalog Data Access and Implementation.

Data in the PGS Catalog is provided under EMBL-EBI's standard terms of use (https://www.ebi.ac.uk/about/terms-of-use/). The data in the Catalog can be currently accessed in the following three ways:

- **Bulk download** of the entire PGS Catalog's metadata, describing all PGS in terms of their publication source, samples used for development/evaluation, and related performance metrics (details and links: www.pgscatalog.org/downloads/).
- The **PGS Catalog FTP server** (available at: https://ftp.ebi.ac.uk/pub/databases/spot/pgs/) is indexed by Polygenic Score (PGS) ID to allow programmatic access to the Scoring Files and metadata for each PGS, archived versions of the scoring files and metadata are also stored for reference (additional details: www.pgscatalog.org/downloads/).
- A **REST API** is also provided to allow programmatic access and querying of the PGS Catalog, better enabling other applications to be built on top of the resource. Endpoints to retrieve all or individual PGS Catalog data objects (Publications, Scores, Samples, Traits, Performance Metrics) are available (details at: https://www.pgscatalog.org/rest/).

The PGS Catalog is also is indexed on FAIRsharing.org (ref: bsg-d001448), and polygenic score identifiers (e.g. PGS000018) can be externally resolved via IDENTIFIERS.org (ref: pgs). A description of the FAIR indicators for the PGS Catalog are provided in Supplemental Table 1.

Additional bibliographic information for PGS Catalog **Publication** objects are retrieved from EuropePMC (e.g. title, authors, journal, publication dates)[1]. Additional information for each ontology term (e.g. synonyms, parent/child relationships, and mapped terms from other ontologies and disease coding resources [e.g. ICD/READ/SNOMED]) from the EFO [2] are obtained using the EMBL-EBI Ontology Lookup Service (OLS)[3].

The PGS Catalog website and database are developed using the Django framework (version 3.1; https://djangoproject.com) in Python (version 3.8; https://www.python.org) with a PostgreSQL database (version 12; https://www.postgresql.org/). The search functionality is built using ElasticSearch (v7.8; https://www.elastic.co).The website, database, and search index are all deployed on the Google Cloud (https://cloud.google.com/). The codebase for the Catalog can be viewed within our public GitHub repository (https://github.com/PGScatalog), currently provided under an Apache 2.0 License.

Supplementary Note 5. Colorectal cancer benchmarking methods.

To evaluate the predictive ability of PGS for colorectal cancer in the Catalog we used data from the UK Biobank (UKB), a cohort of ~500,000 participants from three countries (England, Wales, Scotland) of the United Kingdom[4]. Our analysis included 421,332 participants with genetic and phenotypic data (Supplemental Table 3), corresponding to 409,253 participants of European ancestry (UKB "White British" subset), 6,086 South Asian ancestry, and 5,984 African ancestry participants. South Asian (self-identifying as: Indian, Pakistani, or Bangladeshi) and African ancestry (self-identifying as: Caribbean, African, or Any other black background) participants were defined using an identical process to the White British participants, using principal components of genetic ancestry to identify a homogenous subset of self-identifying individuals by clustering[4].

Diagnosis of colorectal cancer was performed using data linkage to the UK's national cancer and death registries. Cases of colorectal cancer were identified using previously used ICD codes in UKB [5] :

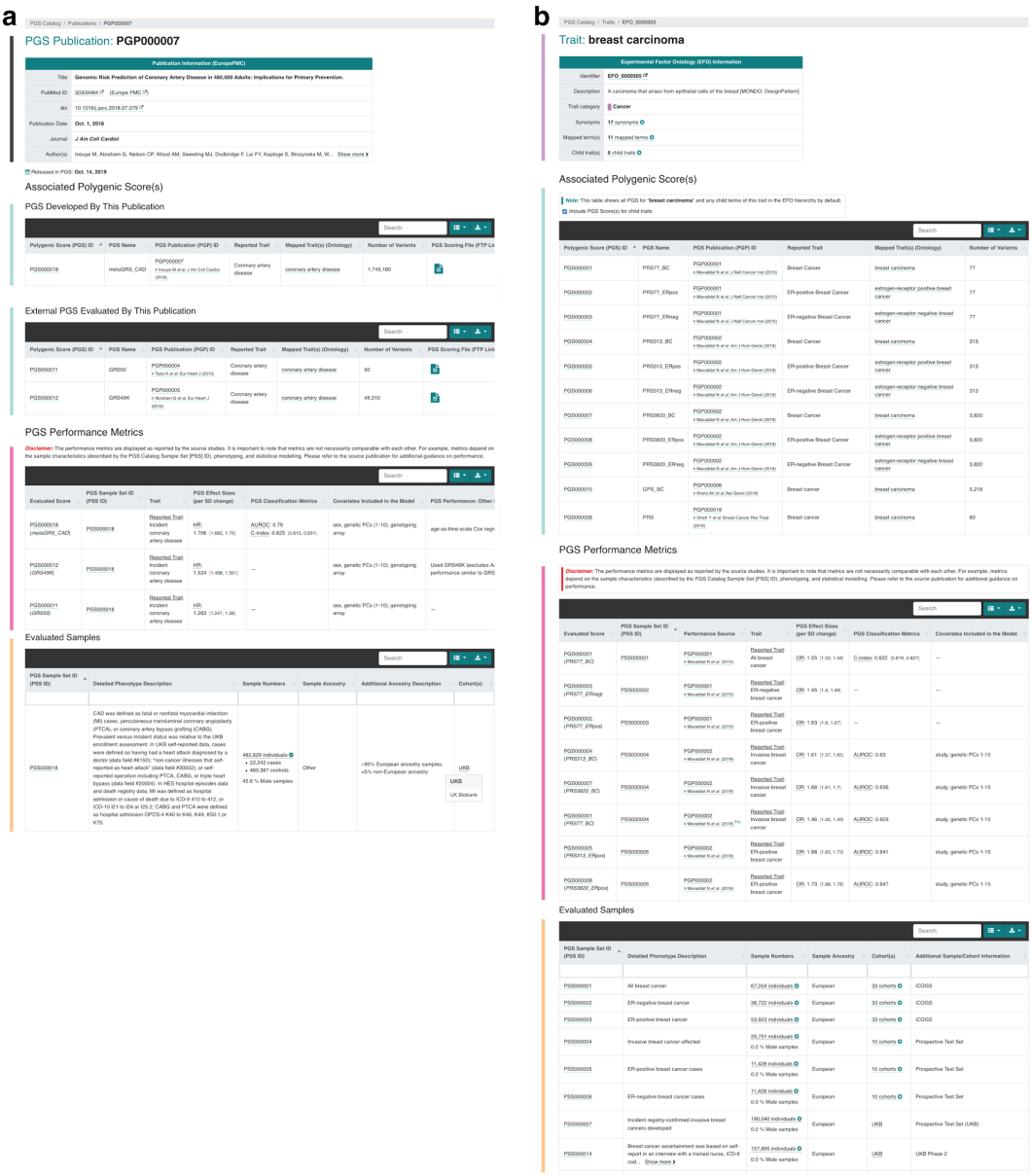
ICD9: 153.0 - 153.9, 154.0, 154.1, 154.8
ICD10: C18.0 - C18.9, C19, C20, C21.8

For each colorectal cancer diagnosis or death we recorded the date and age of the event. colorectal cancer events were defined as the first event of colorectal cancer, and participants were censored after the last cancer registry linkage date (2016-03-31). We excluded 449 participants who had self-reported history of colorectal cancer at recruitment and no linked cancer registry data.

PGS files were downloaded from the PGS Catalog and scores for each participant were calculated using PLINK[6]. Scores were standardised within each ancestry; the mean and standard deviation for colorectal cancer cases and controls are reported by ancestry group (Supplemental Table 3).

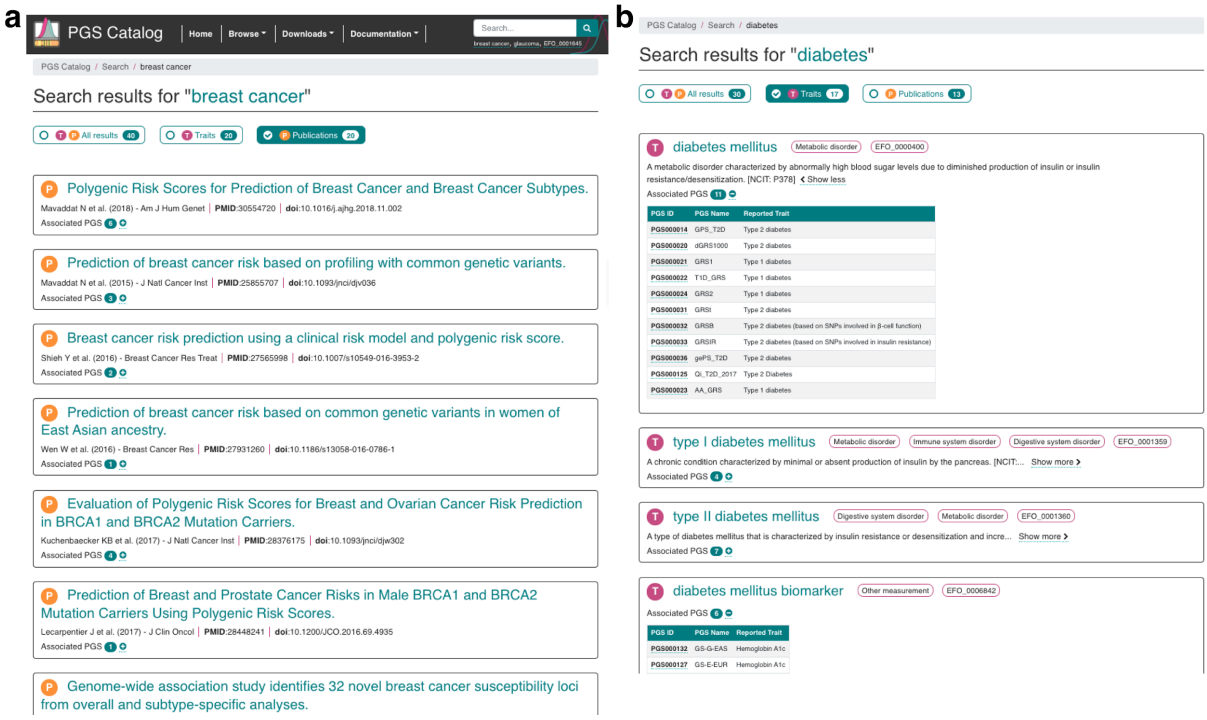
Each score's predictive ability is measured in terms of classification of individuals diagnosed with colorectal cancer versus those without, via the standardised effect size of the PGS (OR/HR per standard deviation increase of PGS) and classification accuracy (AUROC and concordance statistic [C-index]). We measured the HR and C-index using a Cox Proportional Hazards model with age-as-timescale, adjusting for sex, age at recruitment, country of recruitment, genotyping array, and 10 PCs of genetic ancestry. We measured the OR and AUROC using a logistic regression model adjusting for the sex, age at recruitment, country of recruitment, genotyping array, and 10 PCs of genetic ancestry. The effect sizes are reported with the 95% confidence interval for each PGS (Supplemental Table 3). Statistical analyses were performed in python: the Cox model was implemented using the *lifelines* package[7], and logistic regression was performed using the *statsmodels* package[8].

Supplementary Figures



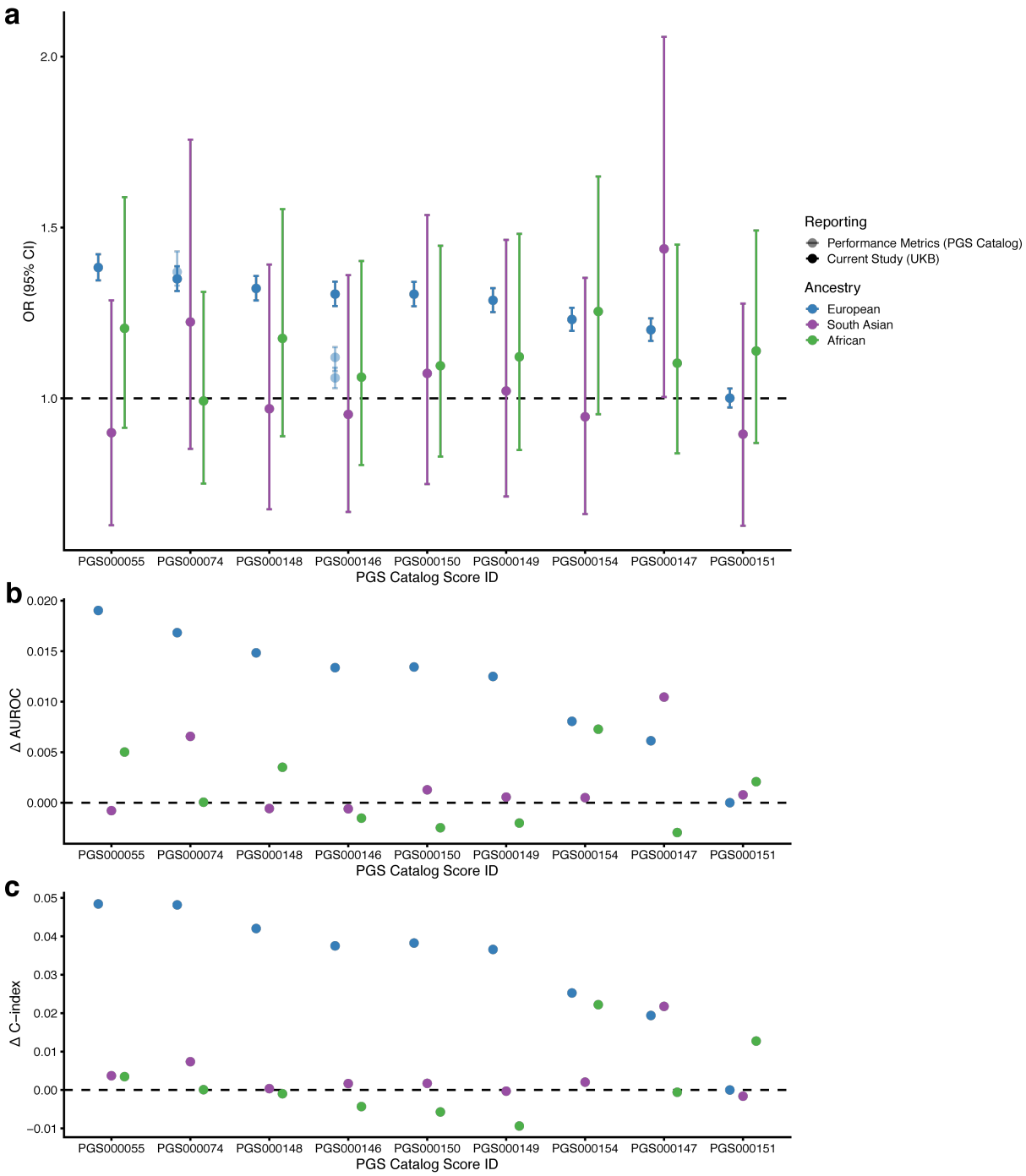
Supplementary Figure 1. Examples of PGS Catalog Publication and Trait website pages.

(**A**) Example of how each Publication and its related metadata (links to publication, EuropePMC, and PGS that were developed and evaluated within the paper) are displayed on PGSCatalog.org (example publication PGP000007[9]). (**B**) Example of how each Trait (ontology term, description, synonyms, mapped terms [e.g. ICD/SNOMED], and child ontology terms/sub-traits extracted from EFO[2,3]) and its related metadata (PGS that have predicted the current trait, and subsequent evaluation of those scores) are displayed on PGSCatalog.org (example trait: breast carcinoma, EFO_0000305). Sub-traits from the ontology (in this example breast cancer subtypes) are displayed by default, but can be removed by de-selecting the “Include PGS Score(s) for child traits” button. Sections of each webpage are highlighted with coloured bars corresponding to the data objects they display in **Figure 1A**.



Supplementary Figure 2. Examples of PGS Catalog search results.

Traits and publications are indexed, and can be queried through the search box at the top right corner of each page's header (**A**). By default the search returns both trait and publication results, but results can be faceted to either. (**A**) Example of search results for publications related to "*breast cancer*". (**B**) Example of search results for traits (ontology terms) related to "*diabetes*". The results include a higher-level "diabetes mellitus" trait (which includes both type 1 and 2), the specific subtypes, and polygenic scores for the related HbA1c measurements under the diabetes mellitus biomarker trait.



Supplementary Figure 3. Performance metrics for colorectal cancer PGS in UKB.

Each PRS was evaluated within a logistic regression model for predicting colorectal cancer status for participants in UKB (**A-B**), and a separate Cox proportional hazards regression model (age-as-timescale) (**Figure 2, C**). (**A**) Standardised effect size (Odds Ratio; OR) describing the odds of having colorectal cancer per unit increase in each PGS. Previously reported effect sizes that were recorded in the Catalog are also plotted for PGS000074 and PGS000146. (**B**) Change in model classification accuracy (Area Under the Receiver Operating Characteristic Curve; ΔAUROC) when the PGS is added to a logistic regression model including the existing covariates (age at recruitment, sex, recruitment country, genotyping array, and 10 PCs of genetic ancestry). (**C**) Change in model classification

accuracy (concordance statistic; ΔC-index) when the PGS is added to a risk model including the existing covariates (sex, age at recruitment, recruitment country, genotyping array, and 10 principal components [PCs] of genetic ancestry).

Supplementary Tables

Supplementary Table 1. FAIR indicators of PGS Catalog.

This table describes details of how the current PGS Catalog conforms to FAIR data principles. For the purposes of this table the Score constitutes the data (e.g. variants, effect weights and alleles), and is linked to metadata (Samples, Performance Metrics, Publications) describing it.

Core FAIR principle	**FAIR principle**	**PGS Catalog indicator**
Findable	F1. (meta)data are assigned a globally unique and persistent identifier	Each polygenic score is assigned a unique identifier (e.g. PGS000018) that is linked to all relevant metadata and publication sources in the Catalog. The PGS identifier can be resolved externally through IDENTIFIERS.org (prefix: *pgs*)
	F2. data are described with rich metadata (defined by R1 below)	Polygenic scores included in the database are well-described, both in terms of their provenance and ability to be applied. Details in Supplemental Table 2 and on our website at: http://www.pgscatalog.org/docs/
	F3. metadata clearly and explicitly include the identifier of the data it describes	All metadata is linked to either a Polygenic Score (PGS), Sample Set (PSS), Performance Metric (PPM), or Publication (PGP) ID within the database. Ontology terms are described using the identifiers from the Experimental Factor Ontology. Publication sources are described using DOI and PMID. Scoring files for each PGS are labelled with their PGS ID, and finable with the metadata on our FTP (http://ftp.ebi.ac.uk/pub/databases/spot/pgs/) described here: http://www.pgscatalog.org/downloads/
	F4. (meta)data are registered or indexed in a searchable resource	The PGS Catalog is indexed at FAIRsharing.org (ID: *bsg-d001448*) and indexed by Google Search.
Accessible	A1. (meta)data are retrievable by their identifier using a standardized communications protocol	Metadata can be easily viewed on our web interface (www.pgscatalog.org) with visible download links for each Score. Scoring files and metadata can also be browsed and downloaded from our FTP site by PGS ID. The full Catalog can also be accessed using our REST API: https://www.pgscatalog.org/rest/.
	A1.1 the protocol is open, free, and universally implementable	Yes, the www.pgscatalog.org website is freely accessible to all.
	A1.2 the protocol allows for an authentication and authorization procedure, where necessary	Not applicable

	A2. metadata are accessible, even when the data are no longer available	Archived versions of the scoring files and metadata are stored for the complete database as well as individual scores on our FTP (http://ftp.ebi.ac.uk/pub/databases/spot/pgs/)
Interoperable	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	PGS metadata is distributed from our API using JSON formats, the REST API is documented using the OpenAPI Specification (OAS3; https://github.com/OAI/OpenAPI-Specification/blob/master/versions/3.0.2.md).
	I2: (Meta)data use vocabularies that follow the FAIR principles	The PGS identifier can be resolved externally through IDENTIFIERS.org (prefix: *pgs*). The traits are consistently described with EFO terms.
	I3. (meta)data include qualified references to other (meta)data	Traits are represented using (represented using ontology terms) associated with PGS are linked to the Experimental Factor Ontology (EFO) terms and include links to the EFO.
Reusable	R1. meta(data) are richly described with a plurality of accurate and relevant attributes	Polygenic scores included in the database are well-described, both in terms of their provenance and ability to be applied. Details in Supplemental Table 2 and on our website at: http://www.pgscatalog.org/docs/
	R1.1. (meta)data are released with a clear and accessible data usage license	All data are made available through EMBL-EBI's standard terms of use (https://www.ebi.ac.uk/about/terms-of-use/). PGS with different licenses and terms are shared openly, but clearly marked with their terms in the metadata and scoring file downloads, as well as beside any download links.
	R1.2. (meta)data are associated with detailed provenance	Each PGS and Performance Metric is linked to a source Publication that can be accessed by either a digital object identifier (DOI) or PubMed ID (PMID).
	R1.3. (meta)data meet domain-relevant community standards	The PGS Catalog is consistent with Polygenic Risk Score Reporting Standards (PRS-RS) [10]

Supplementary Table 2. PGS Catalog reporting items.

This table describes the reporting items that can be captured for each of the data objects in the PGS Catalog.

PGS Catalog Data Objects	Reporting Item	Description	Comments
Publication (Identified by PGP ID)	PubMed ID (PMID)	PubMed Identification number	This information is extracted and annotated according to EuropePMC [1]. Publications are flagged if they are preprints (e.g. *not undergone peer review*).
	Digital Object Identifier (DOI)	The DOI of each publication is curated in addition to the PMID to allow unpublished work (e.g. pre-prints) to be added to the Catalog.	
	Title	Title of the publication or preprint	
	Author(s)	List of publication authors, the first author is also extracted for a shorter display.	
	Journal	The name of the publication source.	
	Publication Date	Date of publication (with respect to the PMID or DOI upon DB upload).	
	Release Date	Date the publication was added to the PGS Catalog.	
Score (Identified by PGS ID)	Reported Trait	The author-reported trait (*e.g.* body mass index [BMI], or coronary artery disease) that the PGS has been developed to predict.	
	Mapped Trait(s)	The Reported Trait is mapped to Experimental Factor Ontology (EFO) terms and their respective identifiers by PGS Catalog curators. For more information about the ontology traits see the Trait object.	Linked to **Ontology Term(s)**.
	PGS Name	This may be the name that the authors use to refer to the PGS, or a name that a curator has assigned to identify the score during the curation process (before a PGS ID has been given).	
	Original Genome Build	The version of the genome that the variants present in the PGS are associated with. Listed as NR (Not Reported) if unknown.	
	Number of Variants	Number of variants used to calculate the PGS. In the future this will include a more detailed description of the types of variants present.	
	Number of Variant Interaction Terms	Number of higher-order variant interactions included in the PGS.	
	PGS Development Method	The name or description of the method or computational algorithm used to develop the PGS.	
	PGS Development Details/Relevant Parameters	A description of the relevant inputs and parameters relevant to the PGS development method/process.	
	Contributing Samples: Source of Variant Associations (GWAS)	Samples used to define the variant associations/effect-sizes used in the PGS. These data are extracted from and linked to the NHGRI-EBI GWAS Catalog when a GWAS study ID (GCST) is provided.	Linked as a **Sample** object(s).

	Contributing Samples: Score Development/Training	Samples used to develop or train the score (*e.g.* not used for variant discovery, and non-overlapping with the samples used to evaluate the PGS predictive ability).	Linked as a **Sample** object(s).
	Publication/Citation	A PGP ID links the PGS to the publication in which it was described.	Linked as a **Publication** object.
	Terms and Licenses	The PGS Catalog distributes its data according to EBI's standard Terms of use. Some PGS have specific terms, licenses, or restrictions (e.g. non-commercial use) that we highlight in this field, if known	
	Release Date	Date the score was added to the PGS Catalog.	
Ontology Term (Mapped traits are identified by an EFO ID)	Name	The trait label from the ontology.	This information is extracted and annotated according to Experimental Factor Ontology (EFO) [2] using the Ontology Lookup Service (OLS) [3].
	Identifier	The Experimental Factor Ontology ID (EFO_ID) identifier to consistently refer to traits using the EFO, and to other resources like the NHGRI-EBI GWAS Catalog.	
	Description	Detailed description of the trait from EFO.	
	Synonyms	Other names for the trait.	
	Mapped Term(s)	Includes references to terms in other databases and ontologies (*e.g.* ICD9/ICD10, MONDO, SNOMEDCT, *etc.*).	
Sample (Groups of samples used in evaluations are given a **Sample Set** [PSS ID])	Number of Individuals	Number of individuals included in the sample	Similar to the GWAS Catalog sample descriptions, and directly extracted from the GWAS Catalog for samples with a GCST ID.
	Number of Cases	Number of individuals with the phenotype of interest (*if dichotomous)*.	
	Number of Controls	Number of individuals without the phenotype of interest (*if dichotomous)*.	
	Percent of participants who are Male	Percent individuals in the sample that are identified as male.	
	Age of Study Participants	A summary of the age distribution(mean/median, range/confidence intervals) of study participants.	
	Broad Ancestral Category	Author reported ancestry is mapped to the best matching ancestry category from the NHGRI-EBI GWAS Catalog framework (Table 1, Morales et al. (2018)).	
	Ancestry	A more detailed description of sample ancestry that usually matches the most specific description described by the authors (e.g. French, Chinese).	
	Country of recruitment	Author reported countries of recruitment (if available).	
	Additional Ancestry Description	Any additional description not captured in the structured data (e.g. founder or genetically isolated populations, or further description of admixed samples).	
	Age of Study Participants	A summary (mean/median, range/confidence intervals) of study participants ages.	
	Participant Follow-up Time	A summary of the follow-up time (mean/median, range/confidence intervals) for participants that are part of a prospective cohort/study design (used to measure disease incidence).	

	Phenotype Definitions and Methods	A description of how the phenotype was measured or defined (e.g. ICD codes used to identify cases/phenotypes in EHR data).	
	Cohort(s)	A list of cohorts that collected the samples.	The initial list of common cohorts used in genetics studies that seeded these annotations is from Mills & Rahal. Communications Biology (2019) [11]
	Additional Sample/Cohort Information	Any additional description about the samples and what they were used for that is not captured by the structured categories (e.g. sub-cohort information).	
Performance Metrics (Identified by a PPM ID)	Evaluated Score		Linked as a **Score** object
	Evaluated Samples	ID that links to the samples the displayed PGS evaluated.	Linked as a **Sample** object(s). Samples used in evaluations are given a **Sample Set** (PSS ID) so that PGS evaluated on the exact same samples can be extracted from the Catalog.
	Trait	This field displays both the Reported and Mapped Traits. The reported trait often corresponds to the test set names reported in the publication, or more specific aspects of the phenotype being tested (e.g. if the disease cases are incident vs. recurrent events).	Can be linked to a **Trait** object.
	Reported Metric: PGS Effect Size	Standardised effect sizes, per standard deviation [SD] change in PGS. Examples include regression coefficients (betas) for continuous traits, Odds ratios (OR) and/or Hazard ratios (HR) for dichotomous traits depending on the availability of time-to-event data.	The reported values of the performance metrics are all reported similarly (e.g. the estimate is recorded along with the 95% confidence interval (if supplied)
	Reported Metric: PGS Classification Metrics	Examples include the Area under the Receiver Operating Characteristic (AUROC) or Harrell's C-index (Concordance statistic).	
	Reported Metric: Other	Metrics that do not fit into the structured categories. Examples include: R2 (proportion of the variance explained), reclassification metrics, p-values from association tests, binned comparisons of PGS risk (e.g. odds ratio of disease risk in the top vs. bottom decile of score).	
	Covariates Included in PGS Model	List of covariates used in the prediction model to evaluate the PGS. Examples include: age, sex, smoking habits, etc.	
	Other Relevant Information	Any other information relevant to the understanding of the performance metrics.	
	Source	ID that links to the publication where the performance metrics were reported.	Linked as a **Publication** object.

Supplementary Table 3. UKB Benchmarking cohort description and results.

Cohort age and sex demographics broken down by colorectal cancer case/control status and participant ancestry. The distribution (mean and standard deviation [SD]) of each standardised PGS in colorectal cancer cases is also given, along with its effect size (Hazard Ratio; HR), citation and number of variants included in the PGS; the distribution of each PGS in controls is zero-mean and unit-variance.

		European		South Asian		African Ancestry	
		Cases	Controls	Cases	Controls	Cases	Controls
Cohort Demographics							
N		5188 (1.28%)	404065	31 (0.51%)	6055	51 (0.86%)	5933
N (Female)		2213	218990	18	2751	30	3503
N (Male)		2975	185075	13	3304	21	2430
Mean age at recruitment (SD)		61.97 (6.15)	57.35 (8.00)	57.87 (7.93)	53.63 (8.45)	58.34 (8.35)	52.87 (8.06)
Mean event/censoring age (SD)		61.47 (8.66)	64.51 (7.98)	58.38 (8.15)	60.43 (8.42)	56.88 (9.96)	59.57 (8.07)
PGS distribution and effect size							
PGS000055		Case PGS Distribution = 0.32 (1.00) HR = 1.38 [1.34 - 1.42]		Case PGS Distribution = -0.10 (0.85) HR = 0.89 [0.63 - 1.28]		Case PGS Distribution = 0.17 (1.07) HR = 1.2 [0.91 - 1.58]	
Schmit SL et al. J Natl Cancer Inst (2019)[12]	76						
PGS000074		Case PGS Distribution = 0.30 (1.01) HR = 1.35 [1.31 - 1.38]		Case PGS Distribution = 0.18 (0.71) HR = 1.21 [0.85 - 1.74]		Case PGS Distribution = -0.02 (0.96) HR = 0.99 [0.75 - 1.31]	
Graff RE et al. bioRxiv (2020)[13]	103						
PGS000146		Case PGS Distribution = 0.26 (1.00) HR = 1.3 [1.27 - 1.34]		Case PGS Distribution = -0.06 (0.88) HR = 0.95 [0.67 - 1.35]		Case PGS Distribution = 0.06 (1.01) HR = 1.06 [0.8 - 1.4]	
Hsu L et al. Gastroenterology (2015)[14]	27						
PGS000147		Case PGS Distribution = 0.18 (1.01) HR = 1.2 [1.17 - 1.23]		Case PGS Distribution = 0.37 (0.97) HR = 1.45 [1.02 - 2.08]		Case PGS Distribution = 0.09 (0.89) HR = 1.1 [0.84 - 1.44]	
Ibáñez-Sanz G et al. Sci Rep (2017)[15]	21						

PGS000148		Case PGS Distribution = 0.28 (1.00) HR = 1.32 [1.28 - 1.35]	Case PGS Distribution = -0.03 (0.84) HR = 0.96 [0.67 - 1.38]	Case PGS Distribution = 0.15 (1.03) HR = 1.17 [0.89 - 1.54]
Jeon J et al. Gastroenterology (2018)[16]	63			
PGS000149		Case PGS Distribution = 0.25 (1.00) HR = 1.28 [1.25 - 1.32]	Case PGS Distribution = 0.01 (0.95) HR = 1.02 [0.71 - 1.47]	Case PGS Distribution = 0.10 (1.10) HR = 1.12 [0.85 - 1.48]
Smith T et al. Br J Cancer (2018)[17]	41			
PGS000150		Case PGS Distribution = 0.26 (1.00) HR = 1.3 [1.27 - 1.34]	Case PGS Distribution = 0.05 (0.91) HR = 1.07 [0.75 - 1.53]	Case PGS Distribution = 0.08 (0.95) HR = 1.09 [0.83 - 1.44]
Weigl K et al. Gastroenterology (2018)[18]	48			
PGS000151		Case PGS Distribution = 0.00 (1.02) HR = 1 [0.97 - 1.03]	Case PGS Distribution = -0.10 (0.88) HR = 0.9 [0.63 - 1.28]	Case PGS Distribution = 0.12 (1.07) HR = 1.13 [0.87 - 1.48]
Xin J et al. Gene (2018)[19]	14			
PGS000154		Case PGS Distribution = 0.20 (1.00) HR = 1.23 [1.2 - 1.26]	Case PGS Distribution = -0.06 (0.93) HR = 0.94 [0.66 - 1.34]	Case PGS Distribution = 0.21 (1.05) HR = 1.25 [0.95 - 1.64]
Shi Z et al. Cancer Med (2019)[20]	30			

Supplementary References

1. Levchenko, M. *et al.* Europe PMC in 2017. *Nucleic Acids Res.* **46,** D1254–D1260 (2018).
2. Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26,** 1112–1118 (2010).
3. Jupp, S., Burdett, T., Leroy, C. & Parkinson, H. E. A new Ontology Lookup Service at EMBL-EBI. in *Proceedings of the 8th Semantic Web Applications and Tools for Life Sciences International Conference, Cambridge UK, December 7-10, 2015* (eds. Malone, J., Stevens, R., Forsberg, K. & Splendiani, A.) **1546,** 118–119 (CEUR-WS.org, 2015).
4. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.

Nature **562,** 203–209 (2018).

5. Saunders, C. L. *et al.* External validation of risk prediction models incorporating common genetic variants for incident colorectal cancer using UK Biobank. *Cancer Prev Res (Phila Pa)* (2020). doi:10.1158/1940-6207.CAPR-19-0521
6. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).
7. Davidson-Pilon, C. lifelines: survival analysis in Python. *JOSS* **4,** 1317 (2019).
8. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in *Proceedings of the 9th Python in Science Conference* 92–96 (SciPy, 2010). doi:10.25080/Majora-92bf1922-011
9. Inouye, M. *et al.* Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J. Am. Coll. Cardiol.* **72,** 1883–1893 (2018).
10. Wand, H. *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *medRxiv* (2020). doi:10.1101/2020.04.23.20077099
11. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2,** 9 (2019).
12. Schmit, S. L. *et al.* Novel common genetic susceptibility loci for colorectal cancer. *J Natl Cancer Inst* **111,** 146–157 (2019).
13. Graff, R. E. *et al.* Cross-Cancer Evaluation of Polygenic Risk Scores for 17 Cancer Types in Two Large Cohorts. *BioRxiv* (2020). doi:10.1101/2020.01.18.911578
14. Hsu, L. *et al.* A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology* **148,** 1330–9.e14 (2015).
15. Ibáñez-Sanz, G. *et al.* Risk Model for Colorectal Cancer in Spanish Population Using Environmental and Genetic Factors: Results from the MCC-Spain study. *Sci. Rep.* **7,** 43263 (2017).
16. Jeon, J. *et al.* Determining risk of colorectal cancer and starting age of screening based on lifestyle, environmental, and genetic factors. *Gastroenterology* **154,** 2152-2164.e19 (2018).

17. Smith, T., Gunter, M. J., Tzoulaki, I. & Muller, D. C. The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the UK Biobank prospective cohort study. *Br. J. Cancer* **119,** 1036–1039 (2018).
18. Weigl, K. *et al.* Genetic risk score is associated with prevalence of advanced neoplasms in a colorectal cancer screening population. *Gastroenterology* **155,** 88-98.e10 (2018).
19. Xin, J. *et al.* Evaluating the effect of multiple genetic risk score models on colorectal cancer risk prediction. *Gene* **673,** 174–180 (2018).
20. Shi, Z. *et al.* Systematic evaluation of cancer-specific genetic risk score for 11 types of cancer in The Cancer Genome Atlas and Electronic Medical Records and Genomics cohorts. *Cancer Med.* **8,** 3196–3205 (2019).