HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2024 June 11.

Published in final edited form as:

Nat Genet. 2021 April ; 53(4): 420–425. doi:10.1038/s41588-021-00783-5.

The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation

Samuel A. Lambert[1,✉,2,3,4], **Laurent Gil**[2,3,5], **Simon Jupp**[4], **Scott C. Ritchie**[1,2,6,7], **Yu Xu**[1,2], **Annalisa Buniello**[4], **Aoife McMahon**[4], **Gad Abraham**[1,8], **Michael Chapman**[2,3,5], **Helen Parkinson**[3,4], **John Danesh**[2,3,5,6,7,9], **Jacqueline A. L. MacArthur**[4,✉], **Michael Inouye**[1,✉,2,3,6,7,8,10]

[1]Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[2]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[3]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK

[4]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK

[5]Wellcome Sanger Institute, Hinxton, UK

[6]National Institute for Health Research Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

[7]British Heart Foundation Cambridge Centre of Research Excellence, Department of Clinical Medicine, University of Cambridge, Cambridge, UK

[8]Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[9]NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[10]The Alan Turing Institute, London, UK

✉ sl925@medschl.cam.ac.uk, jalm@ebi.ac.uk, mi336@medschl.cam.ac.uk.

Author contributions

S.A.L., J.A.L.M. and M.I. conceived the PGS Catalog, led its development and cowrote the manuscript. S.A.L. and L.G. developed the PGS Catalog interface and computational infrastructure, with critical support from S.J. and H.P. S.A.L., J.A.L.M., M.I., A.B., A.M., S.J., G.A., M.C. and J.D. contributed to the definition of relevant PGS metadata. S.A.L., A.B., A.M. and S.C.R. curated data for inclusion in the Catalog. S.A.L. performed the colorectal cancer PGS benchmarking analysis, with contributions from Y.X. and S.C.R. All authors reviewed and contributed edits to the final manuscript.

Competing interests J.D. is part of the International Cardiovascular and Metabolic Advisory Board for Novartis (since 2010); the Steering Committee of UK Biobank (since 2011); the MRC International Advisory Group (ING), London (since 2013); the MRC High Throughput Science Omics Panel, London (since 2013); the Scientific Advisory Committee for Sanofi (since 2013); the International Cardiovascular and Metabolism Research and Development Portfolio Committee for Novartis; and the Astra Zeneca Genomics Advisory Board (2018).

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00783-5.

Peer review information Nature Genetics thanks Melinda Mills and Pradeep Natarajan for their contribution to the peer review of this work.

Abstract

We present the Polygenic Score (PGS) Catalog (https://www.PGSCatalog.org), an open resource of published scores (including variants, alleles and weights) and consistently curated metadata required for reproducibility and independent applications. The PGS Catalog has capabilities for user deposition, expert curation and programmatic access, thus providing the community with a platform for PGS dissemination, research and translation.

By aggregating the effects of many genetic variants into a single number, PGSs have emerged as a method to predict an individual's genetic predisposition to a phenotype[1, 2, 3, 4]. Early studies indicated that combining allelic counts of genome-wide association study (GWAS)-significant variants in individuals is predictive of the phenotype[5, 6, 7, 8]. Owing to larger and more powerful GWAS, recent PGSs typically comprise hundreds to millions of trait-associated genetic variants, which are combined by using a weighted sum of allele doses multiplied by their corresponding effect sizes.

Many PGSs have been developed and demonstrated to be predictive of common complex traits (for example, body mass index (BMI)[9], blood lipids[10] and educational attainment[11]). Similarly, PGSs for various diseases have been shown to be predictive of disease incidence, defining marked increases in risk over the life course or at earlier ages for people with high PGSs (for example, coronary artery disease[12, 13], breast cancer[14] and schizophrenia[15]). Existing risk-prediction models using traditional risk factors can be improved by incorporating PGSs[12, 16, 17]. In some cases, PGSs may be the most informative risk factors in presymptomatic individuals[1, 18] and, for some diseases, may be independent of a family history of the condition[19, 20, 21, 22]. Other potential clinical uses of PGSs include prediction of prognosis, etiology and disease subtypes[23]; stratification of patients according to therapeutic benefit; and identification of new disease biomarkers and drug targets[24]. Given their multiple applications, many PGSs have been developed, and more than 1,000 related articles have been indexed in PubMed since 2009.

There is widespread variability in PGS research, even regarding nomenclature: the scores can be referred to as genetic or genomic scores, and as polygenic risk scores (PRSs) or genomic risk scores (GRS) if they predict a discrete phenotype (such as a disease)[25]. Many approaches also exist to derive PGSs by using individual-level genotype data or GWAS summary statistics[26]. The goals of most computational methods are to select the most predictive set of variants in the score, and to adjust their weights to maximize the predictive ability and account for linkage disequilibrium between variants.

The need for an open resource of polygenic scores

Multiple barriers hinder progress in PGS research and the translation of PGSs into healthcare settings. The lack of best practices and standards, particularly regarding PGS reporting, is a major issue identified by our group and others[25, 27]. Reproducibility has been hampered by underreporting of key PGS information: approximately 40% of 231 publications developing new PGSs that we reviewed during our curation efforts did not include adequate variant information (for example, chromosomal location, effect allele and

weight) to calculate the PGSs for new samples, thus limiting the utility and reusability of the scores.

Beyond the information necessary for PGS calculation, a complete understanding of a score's ability to accurately predict its target trait (also known as analytic validity) is necessary to help evaluate clinical utility and enable other applications of PGSs. However, the performance metrics reported for existing PGSs are conditional on study design, participant demographics, case definitions and the covariates adjusted for in the original studies' models. Although few direct evaluations of PGSs have been conducted, benchmarking of multiple PGSs for the same trait in external data provides directly comparable performance metrics[28] needed to decide which PGS has the best performance for a particular task and how its predictive ability varies in response to changes in important factors, such as ancestry[29]. Because PGSs are based on data and cohorts composed of largely European ancestry individuals, there is a well-characterized underperformance of PGSs when they are applied to non-European ancestry individuals; thus, the transferability of PGS performance is a particularly important challenge that could lead to health disparities[30, 31, 32].

Here, we present the PGS Catalog, an open resource of published PGSs, including full scoring information annotated with expertly curated metadata required for accurate application and evaluation. The PGS Catalog promotes PGS reproducibility by providing a venue to annotate and distribute scores according to current exemplar reporting standards. As such, it allows users to reuse and evaluate PGSs, to firmly establish their predictive ability and facilitate further investigations of clinical utility.

Development of the PGS Catalog

The aim of the PGS Catalog is to index and distribute the key aspects of each PGS (underlying variants, results and experimental design) in a standardized representation, to facilitate evaluation of analytic validity. To maximize usability, the data representation and database were designed to be findable, accessible, interoperable and reusable (FAIR) according to established principles for scientific data management[33] (Supplementary Table 1).

To define the key information that would need to be captured in the PGS Catalog, we undertook an initial literature review of publications that developed PGSs for the following traits and diseases, according to their potential clinical utility and public-health burden of disease: coronary artery disease, diabetes (types 1 and 2), obesity/BMI, breast cancer, prostate cancer and Alzheimer's disease. During our review, we took note of how the PGSs were described, how they differed between studies and traits, and the most common study designs and PGS evaluation scenarios. To capture common aspects of PGS studies, we built upon the NHGRI-EBI GWAS Catalog's established frameworks to catalog published data from genomic studies, by using accepted conventions for representing sample ancestry[34], variant and trait information[35]. Using our survey and established frameworks, we defined four major data objects: scores, samples, performance metrics and publications (Box 1 and Supplementary Table 2). These objects describe the common PGS development and

evaluation processes (Fig. 1a), and can be used to capture the detailed data elements necessary to evaluate PGS development and performance. High consistency and accuracy across curated data were ensured by developing detailed curation guidelines, inclusion criteria and data acquisition methods (outlined in the Supplementary Note).

To ensure that the PGS Catalog contains the information necessary to describe and evaluate PGSs, we collaborated with the ClinGen Complex Disease working group, composed of experts in epidemiology, statistics, implementation science and the actionability of genetic results, as well as those with disease-domain-specific knowledge and interests in PRS application. Together, we developed the Polygenic Risk Score Reporting Standards (PRS-RS)[25], a joint statement describing a set of reporting items that should be described in studies developing and evaluating PRSs. The PGS Catalog captures the data required by the PRS-RS to assess PGS validity while also being flexible enough to capture multiple different study designs and evaluation scenarios in a structured database. The PGS Catalog therefore provides a venue to index PGS analyses and maximize uptake of these reporting standards.

Scores (for example, PGSs, PRSs or GRSs) are the main data object type in the PGS Catalog, are linked to all other objects internally, and can be cited or externally linked to through a persistent identifier (for example, PGS000018). Each PGS has a PGS scoring file—a flat text file in a consistent format (Supplementary Note), which contains the variant-level information necessary to calculate the score on new data (minimally the genome build, rsID or chromosomal positions, effect alleles and their weights). The PGS is also annotated with information about the phenotype that it predicts (reported trait) and is mapped to Experimental Factor Ontology terms[41, 42] to consistently annotate related scores and facilitate data linkage and searching. Information describing the computational algorithms (for example, independent GWAS variants, pruning/clumping and thresholding, or LDpred) and parameters (for example, *P*-value and linkage-disequilibrium (r^2) thresholds) used during score development are also recorded for each score. The GWAS summary statistics used to derive the PGS, if any, are linked as sample objects and further linked to the GWAS Catalog if applicable[35], and any other datasets used for training are also linked as sample objects.

Samples are described with detailed information to enable the interpretation and assessment of the validity of a PGS. Sample size (stratified by cases and controls if dichotomous) and participant ancestry are described by using frameworks identical to those in the GWAS Catalog[34] to enable the systematic tracking of participant diversity in PGSs[32]. To facilitate reproducible analyses, phenotyping descriptions (for example, case definition, International Classification of Diseases 9/10 codes, and measurement methods), the sex distribution, and the distributions of participant ages and follow-up times for prospective study designs can also be recorded. To ensure that PGSs are not evaluated on individuals who contributed to the original GWAS or PGS training cohorts, samples can be annotated with existing cohort names[43]. Groups of samples used to evaluate a PGS are given a Sample S et ID Sample Set is sort of like a proper noun, like a named object class.

Performance metrics assess the validity of a PGS in a Sample Set independently of the samples used for score development. Common metrics include standardized effect sizes

(odds ratios or hazard ratios, and regression coefficients (β)), classification accuracy metrics (for example, area under the receiver operating characteristic curve, C-index and area under the precision-recall curve), but other relevant metrics (for example, calibration (χ^2)) can also be recorded. The covariates used in the model (most commonly age, sex and genetic principal components to account for the population structure) are also recorded for each set of metrics. Multiple PGSs can be evaluated on the same sample set and further indexed as directly comparable performance metrics.

Publications provide provenance information for scores and performance metrics (including those from external evaluations of existing PGSs). Both journal articles and preprints can be indexed through either the doi number or PubMed ID.

The PGS Catalog: data content, access and expansion

Any published or preprinted PGS can be added to the PGS Catalog, provided that it has (1) established analytic validity in external samples not used for score development and (2) the information necessary to calculate the score (additional details in Supplementary Note). To populate the PGS Catalog, we screened more than 275 publications for eligibility, 162 of which presented sufficient data for curation and inclusion in the Catalog. As of December 2020, the PGS Catalog contains 657 consistently annotated PGSs curated from 119 publications (with the earliest published in 2008). These PGSs predict a wide variety of diseases (for example, cardiovascular diseases, different types of cancer, schizophrenia, major depressive disorder), as well as anatomical (for example, BMI and bone density), cellular (for example, blood cell phenotypes and counts) and molecular (for example, serum urate, cholesterol and triglyceride levels) traits and measurements, encompassing 156 unique mapped ontology terms. Currently, most PGSs included in the Catalog were developed in European ancestry individuals; however, 11 PGSs were both developed and evaluated in individuals of non-European ancestry. To assess external validity, the Catalog also indexes the results of evaluations of existing PGSs in new contexts (for example, direct comparisons of multiple PGSs on the same sample); 13 of these benchmarking publications evaluating 14 existing PGSs are also included in the current release of the PGS Catalog. Of the 119 publications, 105 developed at least one new PGS, of which 14 also included a benchmarking of the performance to existing PGSs.

The PGS Catalog can be accessed through an online user interface The main URL (https://www.PGSCatalog.org) should remain/appear in this section. (https://www.PGSCatalog.org) in which indexed publications, scores and traits are browsable and searchable. Metadata describing PGS development and evaluation can be viewed on each score's page (annotated example in Fig. 1b). Pages describing traits with available PGSs, and the scores developed and evaluated within each publication can also be viewed (Supplementary Fig. 1). Trait pages display any PGSs associated with subtraits by default (for example, scores for all breast cancer subtypes are displayed on the breast cancer trait page), and higher-level disease and trait categories are accessible via our ontology-enriched search (Supplementary Fig. 2). Links to relevant study pages in the GWAS Catalog are included for any score developed by using cataloged GWAS data. Navigation from the GWAS Catalog to relevant

data in the PGS Catalog is supported through links on the publication, study and trait pages in the GWAS Catalog.

Each PGS in the Catalog is provided as a scoring file containing a header describing the provenance of the score, and consistently formatted columns describing the variants, alleles and weights. The scoring file can be used in conjunction with common tools to calculate the PGS (for example, PLINK[36]). The metadata and scoring files can be downloaded alone or in bulk from our website and FTP server; programmatic access to the database is also available through a RESTful API (complete implementation and scoring file details in the Supplementary Note). Importantly, the PGS Catalog provides users with a source of existing published PGSs that can be directly applied to their own data, thus making results obtained by using the same score comparable across users and use cases, and circumventing the need to develop a new PGS for every application.

The Catalog identifies new articles from a manual literature search and user submissions, which subsequently undergo curation before their inclusion (Supplementary Note). Data curation and submission have been designed around a flexible template that allows common PGS development and evaluation details and results to be described according to our reporting items, and the template and PGS can be submitted directly to the Catalog for inclusion after validation by curators. Authors of PGS studies are encouraged to submit new PGSs as well as the results of subsequent PGS validations for indexing (by e-mail to pgs-info@ebi.ac.uk; more information at https://www.PGSCatalog.org/submit), to grow the Catalog for the community, maximize the utility of their PGSs and support reproducibility.

Generating comparable PGS performance metrics

Where multiple PGSs have been cataloged for a trait of interest, a complete understanding of the predictive ability of each PGS would be useful for deciding which score is best for a user's particular application. However, the performance metrics of PGSs are not directly comparable (owing to differences in samples or cohorts, covariates and study design) and have usually been measured in only a single ancestry group. To demonstrate how the PGS Catalog can be used to systematically compare PGS performance, we measured the performance of nine PGSs for colorectal cancer in people of European, South Asian and African ancestries in the UK Biobank (UKB)[37], a dataset external to the development of all scores (methods described in Supplementary Note; cohort described in Supplementary Table 3). For each ancestry group, each PGS was evaluated by using the standardized effect size of the PGS (odds ratio/hazard ratio per s.d. increase in PGS) and changes in classification accuracy (area under the receiver operating characteristic curve and C-index) as performance metrics (Fig. 2 and Supplementary Fig. 3). Eight of the nine scores were predictive of colorectal cancer in European ancestries in the UKB to varying degrees, and the magnitudes of the effect sizes for two PGSs were similar to that previously reported (Supplementary Fig. 3). The score not significantly predictive of colorectal cancer in European ancestry participants (PGS000151) comprised only 14 variants, and its predictive ability in Europeans had not previously been evaluated. Although the majority of scores were predictive of colorectal cancer in the 409,253 European ancestry participants, the PGSs were largely not predictive of disease risk in the 6,086 participants of South Asian ancestry

and 5,984 participants of African ancestry (together composing approximately 8% of all UKB participants; Supplementary Table 3); these data further illustrate that some PGSs developed by using European-biased GWAS data have lesser predictive ability and may not be valid in people of non-European ancestry[30, 32]. The PGS Catalog will continue to curate and generate PGS benchmarking data from participants with diverse ancestries to provide a more comprehensive understanding of PGS performance, which is necessary to prioritize the best PGS for a particular application.

Conclusions and future developments

The PGS Catalog is a publicly available resource of published PGSs. The Catalog makes PGSs for any heritable trait available for analysis in a standardized format along with consistent metadata, thereby enabling direct comparison between scores. In these ways, the Catalog differs from previous databases that are disease specific (for example, Cancer PRSweb[38]) or those that distribute phenome-wide PRS associations for bespoke scores [39], and it serves the community by providing a platform for PGS distribution and research.

We hope to facilitate reproducible PGS analyses by working with others toward developing standard formats and content of scoring files, and providing new tools (such as for validation and scoring) to support this aim. For instance, to address a common user request, we will harmonize variants in PGS scoring files to frequently used genome builds (GRCh37 and 38). PGS reproducibility must also ensure that calculations are valid and consistent, with minimal variability across users. According to community need, we intend to provide reference-sample calculations and population distributions similar to those for clinical tests. These enhancements will facilitate systematic and external PGS benchmarking studies, which are key in evaluating the validity of existing PGSs.

As PGSs increase in number, and the diversity of the phenotypes that they predict increases, we will continue to grow the Catalog. We will be developing an interface to support author submission of developed and evaluated PGSs, providing an accession ID at the point of submission to enable citation. We are also working toward an improved literature search to reliably identify published PGSs. This search will be used to provide a regularly updated comprehensive index of PGS publications in the Catalog, with a page for each publication regardless of data availability; a more systematic resource will better highlight gaps in the field (for example, a lack of PGSs developed for participants of non-European ancestry) [32]. Our evaluation of PGS performance across ancestries further underscores the need to develop, share and evaluate PGSs in diverse populations to avoid subsequent health disparities and inaccurate interpretations from studies using PGSs to evaluate differences among populations [30, 40]. Methodological improvements and more diverse participation in biobanks are likely to overcome these limitations and ideally empower future applications of PGSs in research and clinical settings.

While curating publications for the Catalog, we encountered numerous barriers to making PGS data available. These included restrictions on sharing the variants and weights in a PGS, owing to commercial interests, the need to accept terms and conditions to access the underlying GWAS summary statistics, or researchers not sharing the PGS and indicating

that the full summary statistics were sufficient even if the PGS was based on filtered or reweighted variants. We hope that researchers developing PGSs in the future will consider the need to share their data and PGSs, and adopt reporting standards (PRS-RS[25]) to enable reproducibility as well as subsequent applications and translation of the scores that they have developed. We encourage all researchers, funders and publishers to promote data sharing and submission so that the PGS Catalog can provide a comprehensive resource for the community.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We wish to thank all the authors of publications in the PGS Catalog for making their data available and indexable in our database, and all those who responded to our inquiries and requests for data. We thank P. L. Whetzel for implementing the links from the NHGRI-EBI GWAS Catalog publication, study and trait pages to the PGS Catalog. We also wish to acknowledge E. Tinsley, S. Saverimuttu and members of the laboratory of M.I. for curation support. This work makes use of data from UK Biobank Project no. 7439.

This work was supported by core funding from the UK Medical Research Council (MR/L003120/1), the British Heart Foundation (RG/13/13/30194 and RG/18/13/33946) and the National Institute for Health Research (Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust). This work was also supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. M.I. was supported by the Munz Chair of Cardiovascular Prediction and Prevention. This study was supported by the Victorian Government's Operational Infrastructure Support (OIS) program. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award U41HG007823. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. In addition, we acknowledge funding from the European Molecular Biology Laboratory. J.D. holds a British Heart Foundation Professorship and is funded by a National Institute for Health Research Senior Investigator Award. M.I. and S.R. are supported by the National Institute for Health Research (Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust). S.A.L. is supported by a Canadian Institutes of Health Research postdoctoral fellowship (MFE-171279). This work was performed by using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (http://www.csd3.cam.ac.uk), provided by Dell EMC and Intel, by using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1) and DiRAC funding from the Science and Technology Facilities Council (http://www.dirac.ac.uk). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

References

1. Lambert SA, Abraham G & Inouye M Hum. Mol. Genet. 28, R133–R142 (2019). (R2). [PubMed: 31363735]
2. Torkamani A, Wineinger NE & Topol EJ Nat. Rev. Genet. 19, 581–590 (2018). [PubMed: 29789686]
3. Chatterjee N, Shi J & García-Closas M Nat. Rev. Genet. 17, 392–406 (2016). [PubMed: 27140283]
4. Wray NR, Kemper KE, Hayes BJ, Goddard ME & Visscher PM Genetics 211, 1131–1141 (2019). [PubMed: 30967442]
5. Evans DM, Visscher PM & Wray NR Hum. Mol. Genet. 18, 3525–3531 (2009). [PubMed: 19553258]
6. International Schizophrenia Consortium. Nature 460, 748–752 (2009). [PubMed: 19571811]
7. Kathiresan S et al. N. Engl. J. Med. 358, 1240–1249 (2008). [PubMed: 18354102]
8. Ripatti S et al. Lancet 376, 1393–1400 (2010). [PubMed: 20971364]

9. Khera AV et al. Cell 177, 587–596.e9 (2019). [PubMed: 31002795]
10. Kuchenbaecker K et al. Nat. Commun. 10, 4330 (2019). [PubMed: 31551420]
11. Lee JJ et al. Nat. Genet. 50, 1112–1121 (2018). [PubMed: 30038396]
12. Inouye M et al. J. Am. Coll. Cardiol. 72, 1883–1893 (2018). [PubMed: 30309464]
13. Khera AV et al. Nat. Genet. 50, 1219–1224 (2018). [PubMed: 30104762]
14. Mavaddat N et al. Am. J. Hum. Genet. 104, 21–34 (2019). [PubMed: 30554720]
15. Zheutlin AB et al. Am. J. Psychiatry 176, 846–855 (2019). [PubMed: 31416338]
16. Abraham G et al. Nat. Commun. 10, 5819 (2019). [PubMed: 31862893]
17. Mavaddat N et al. J. Natl. Cancer Inst. 107, djv03 (2015).
18. Natarajan PJ Am. Coll. Cardiol. 72, 1894–1897 (2018).
19. Tada H et al. Eur. Heart J. 37, 561–567 (2016). [PubMed: 26392438]
20. Abraham G et al. Eur. Heart J. 37, 3267–3278 (2016). [PubMed: 27655226]
21. Seibert TM et al. Br. Med. J. 360, j5757 (2018). [PubMed: 29321194]
22. Li H et al. Genet. Med. 19, 30–35 (2017). [PubMed: 27171545]
23. Sharp SA et al. Diabetes Care 42, 200–207 (2019). [PubMed: 30655379]
24. Natarajan P et al. Circulation 135, 2091–2101 (2017). [PubMed: 28223407]
25. Wand H et al. Nature https://doi.org/10.1038/sxxx (2021).
26. Choi SW, Mak TS-H & O'Reilly PF Nat. Protoc. 15, 2759–2772 (2020). [PubMed: 32709988]
27. ICDA Organizing Committee and Working Groups. International Common Disease Alliance Recommendations and White Paper v.1.0 (ICDA, 2020); https://drive.google.com/file/d/16SVJ5lbneN9hB9E03PZMhpescAN527HO/view.
28. Wünnemann F et al. Circ. Genom. Precis. Med. 12, e002481 (2019). [PubMed: 31184202]
29. Dikilitas O et al. Am. J. Hum. Genet. 106, 707–716 (2020). [PubMed: 32386537]
30. Martin AR et al. Nat. Genet. 51, 584–591 (2019). [PubMed: 30926966]
31. Reisberg S, Iljasenko T, Läll K, Fischer K & Vilo J PLoS ONE 12, e0179238 (2017). [PubMed: 28678847]
32. Duncan L et al. Nat. Commun. 10, 3328 (2019). [PubMed: 31346163]
33. Wilkinson MD et al. Sci. Data 3, 160018 (2016). [PubMed: 26978244]
34. Morales J et al. Genome Biol. 19, 21 (2018). [PubMed: 29448949]
35. Buniello A et al. Nucleic Acids Res. 47, D1005–D1012 (2019). (D1). [PubMed: 30445434]
36. Chang CC et al. Gigascience 4, 7 (2015). [PubMed: 25722852]
37. Bycroft C et al. Nature 562, 203–209 (2018). [PubMed: 30305743]
38. Fritsche LG et al. Am. J. Hum. Genet. 107, 815–836 (2020). [PubMed: 32991828]
39. Richardson TG, Harrison S, Hemani G & Davey Smith G eLife 8, e43657 (2019). [PubMed: 30835202]
40. Rosenberg NA, Edge MD, Pritchard JK & Feldman MW Evol. Med. Public Health 2019, 26–34 (2018). [PubMed: 30838127]
41. Malone J et al. Bioinformatics 26, 1112–1118 (2010). [PubMed: 20200009]
42. Jupp S, Burdett T, Leroy C & Parkinson HE In Proc. 8th International Conference on Semantic Web Applications and Tools for Life Sciences (eds. Malone J. et al.) 1546, 118–119 (CEUR Workshop Proceedings, 2015).
43. Mills MC & Rahal C Commun. Biol. 2, 9 (2019). [PubMed: 30623105]

Box 1

Descriptions of PGS Catalog objects and metadata

Individual reporting items are described field by field in Supplementary Table 2. Could this be displayed in a subtitle style of text (perhaps in italics)? It currently looks like regular text but it's meant to be a subtitle/pointer to relevant information in the supplement.

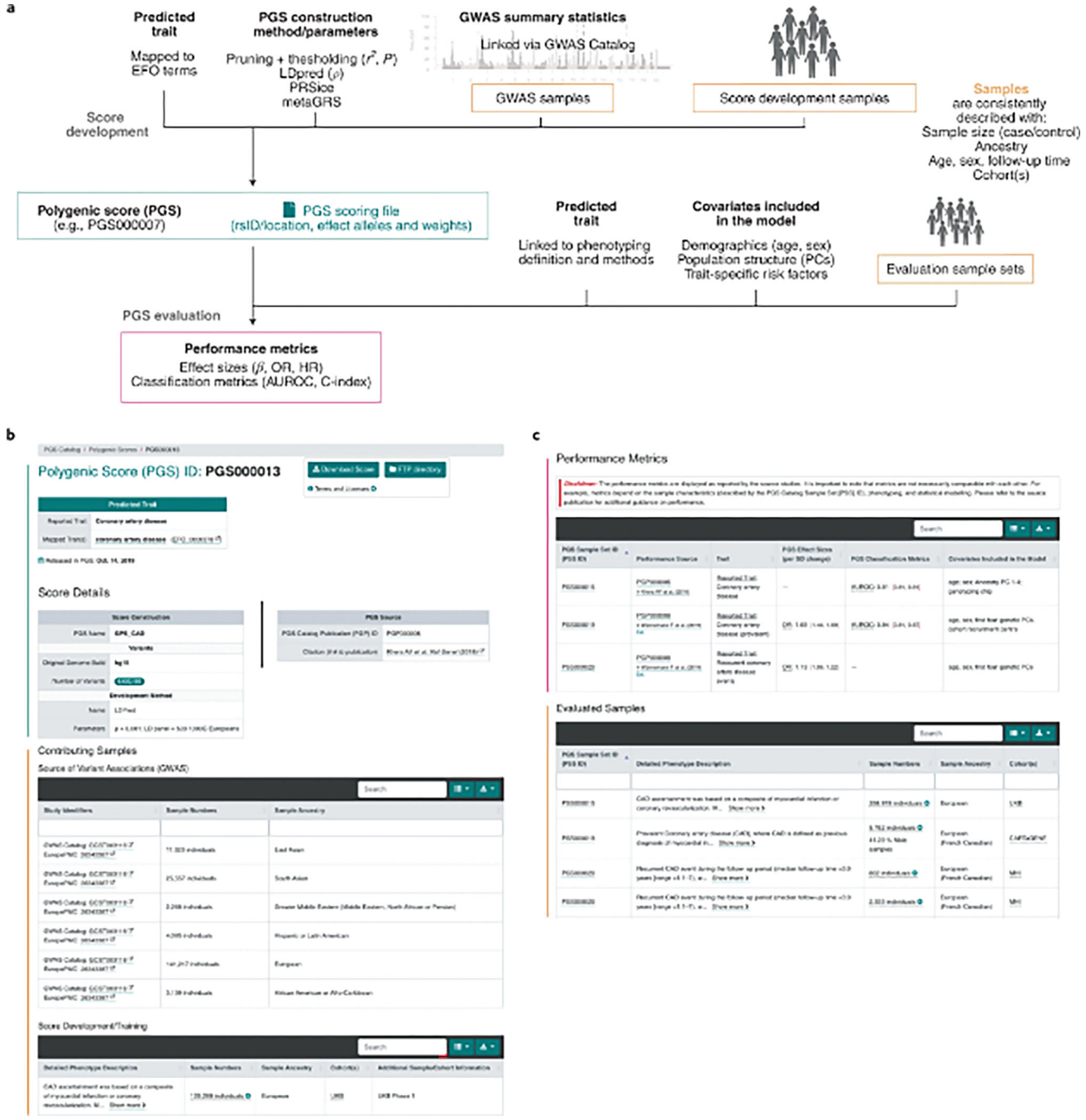

Fig. 1. Common aspects of PGS analyses that are captured and displayed in the PGS Catalog.

a, PGS analyses can broadly be described in two stages: determining the set of variants and weights that will predict a trait of interest (score development) and an evaluation of how predictive the PGS is in an external set of samples (PGS evaluation). Major data items (Box 1) that can be queried and browsed in the PGS Catalog are highlighted as colored boxes and linked to metadata items that are recorded. **b**,**c**, Examples of how PGS metadata are displayed for each score on https://www.PGSCatalog.org (example score PGS000013; ref. [13]), including score details, contributing samples and score development/training (**b**) and

performance metrics and evaluated samples (**c**). Sections are highlighted with colored bars corresponding to the data objects that they display in **a**.

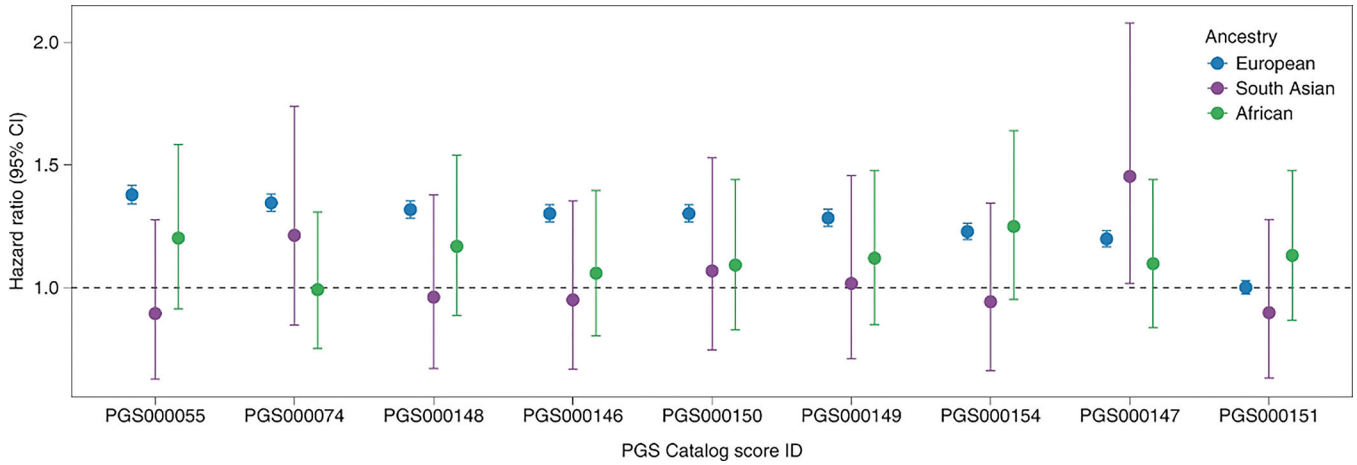

Fig. 2. Benchmarking the association of nine colorectal cancer PGSs in UKB.

Each PGS was evaluated with a Cox proportional hazards regression model (age as timescale) to predict colorectal cancer status. Each model was fitted separately for each ancestry group. The standardized effect size (hazard ratio), together with the 95% confidence interval (CI), describes the increase in hazard per s.d. increase in each PGS. Models were adjusted for sex, age at recruitment, recruitment country, genotyping array and the first ten genetic principal components within each ancestry group.