



# PGS Catalog access with quincunx

## Introduction

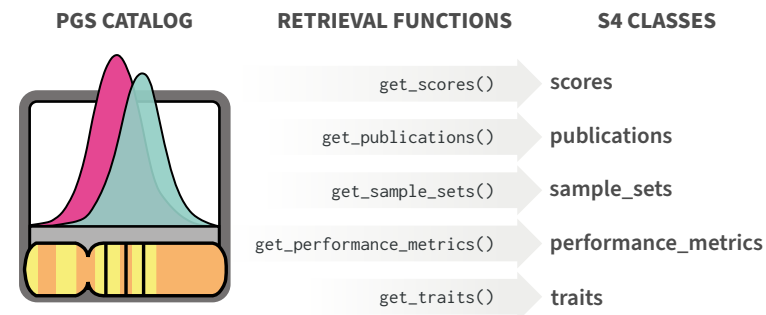
The **PGS Catalog** is a service provided by the EMBL-EBI and NHGRI that offers a manually curated and freely available database of published polygenic scores (PGS): <https://www.pgscatalog.org/>.

The PGS Catalog data provided by the **RESTful API** is organised around five core entities:

- PGS** Polygenic Scores
- PGP** PGS Publications
- PSS** PGS Sample Sets
- PPM** PGS Performance Metrics
- EFO** EFO traits

## Get PGS Catalog Entities

**quincunx** facilitates the access to the Catalog via the RESTful API, allowing you to programmatically retrieve data directly into R. Each of the five entities is mapped to an S4 object of a class of the same name.



Query criteria for retrieval functions, e.g., PGS can be queried by either `pgs_id`, `efo_id` or `pubmed_id`. These correspond to the criteria exposed by the PGS Catalog REST API: <https://www.pgscatalog.org/rest/>.

| Search by  | Example       | PGS | PGP | PSS | PPM | EFO |
|------------|---------------|-----|-----|-----|-----|-----|
| pgs_id     | "PGS000001"   |     |     |     |     |     |
| pgp_id     | "PGP000001"   |     |     |     |     |     |
| pss_id     | "PSS000001"   |     |     |     |     |     |
| ppm_id     | "PPM000001"   |     |     |     |     |     |
| efo_id     | "EFO_0000249" |     |     |     |     |     |
| pubmed_id  | "25855707"    |     |     |     |     |     |
| author     | "Mavaddat"    |     |     |     |     |     |
| trait_term | "Alzheimer"   |     |     |     |     |     |

## PGS Catalog Entities in R

PGS Catalog entities are represented as S4 classes in R. Each class represents a relational database of tidy data tables. All objects start with a table with the same name as the class. Combination of variables indicated in bold renders each row unique in each table.

### S4 class scores

|   |   |   |
|---|---|---|
| <b>scores</b> <ul style="list-style-type: none"><li><b>pgs_id</b></li><li>pgs_name</li><li>scoring_file</li><li>matches_publication</li><li>reported_trait</li><li>trait_additional_description</li><li>pgs_method_name</li><li>pgs_method_params</li><li>n_variants</li><li>n_variants_interactions</li><li>assembly</li><li>license</li><li>beta_unit</li></ul> | <b>samples</b> <ul style="list-style-type: none"><li><b>pgs_id</b></li><li><b>sample_id</b></li><li>stage</li><li>sample_size</li><li>sample_cases</li><li>sample_controls</li><li>sample_percent_male</li><li>phenotype_description</li><li>ancestry</li><li>ancestry_description</li><li>ancestry_country</li><li>ancestry_additional_description</li><li>study_id</li><li>pubmed_id</li><li>cohorts_additional_description</li></ul> | <b>demographics</b> <ul style="list-style-type: none"><li><b>pgs_id</b></li><li><b>sample_id</b></li><li><b>variable</b></li><li>estimate_type</li><li>estimate</li><li>unit</li><li>variability_type</li><li>variability</li><li>interval_type</li><li>interval_lower</li><li>interval_upper</li></ul> |
| <b>publications</b> <ul style="list-style-type: none"><li><b>pgs_id</b></li><li><b>pgp_id</b></li><li>pubmed_id</li><li>publication_date</li><li>publication</li><li>title</li><li>author_fullname</li><li>doi</li></ul>  | <b>traits</b> <ul style="list-style-type: none"><li><b>pgs_id</b></li><li><b>efo_id</b></li><li>trait</li><li>description</li><li>url</li></ul>   | <b>cohorts</b> <ul style="list-style-type: none"><li><b>pgs_id</b></li><li><b>sample_id</b></li><li><b>cohort_symbol</b></li><li>cohort_name</li></ul>  |

### S4 class publications

|  |  |
|--|--|
| <b>publications</b> <ul style="list-style-type: none"><li><b>pgp_id</b></li><li>pubmed_id</li><li>publication_date</li><li>publication</li></ul> | <b>pgs_ids</b> <ul style="list-style-type: none"><li><b>pgp_id</b></li><li><b>pgs_id</b></li><li>stage</li></ul> |
|--|--|

### S4 class traits

|   |   |   |
|---|---|---|
| <b>traits</b> <ul style="list-style-type: none"><li><b>efo_id</b></li><li><b>parent_efo_id</b></li><li>is_child</li><li>trait</li><li>description</li><li>url</li></ul>   | <b>pgs_ids</b> <ul style="list-style-type: none"><li><b>efo_id</b></li><li><b>parent_efo_id</b></li><li>is_child</li><li>pgs_id</li></ul> | <b>child_pgs_ids</b> <ul style="list-style-type: none"><li><b>efo_id</b></li><li><b>parent_efo_id</b></li><li>is_child</li><li>child_pgs_id</li></ul> |
| 3x <b>trait</b> {categories, synonyms, mapped_terms} <ul style="list-style-type: none"><li><b>efo_id</b></li><li><b>parent_efo_id</b></li><li>is_child</li><li>trait_{category, synonyms, mapped_terms}</li></ul> |   |   |

### S4 class sample\_sets

|  |   |   |
|--|---|---|
| <b>sample_sets</b> <ul style="list-style-type: none"><li><b>pss_id</b></li><li>pgs_name</li><li>scoring_file</li><li>matches_publication</li><li>reported_trait</li><li>trait_additional_description</li><li>pgs_method_name</li><li>pgs_method_params</li><li>n_variants</li><li>n_variants_interactions</li><li>assembly</li><li>license</li><li>beta_unit</li></ul> | <b>samples</b> <ul style="list-style-type: none"><li><b>pss_id</b></li><li><b>sample_id</b></li><li>stage</li><li>sample_size</li><li>sample_cases</li><li>sample_controls</li><li>sample_percent_male</li><li>phenotype_description</li><li>ancestry</li><li>ancestry_description</li><li>ancestry_country</li><li>ancestry_additional_description</li><li>study_id</li><li>pubmed_id</li><li>cohorts_additional_description</li></ul> | <b>demographics</b> <ul style="list-style-type: none"><li><b>pss_id</b></li><li><b>sample_id</b></li><li><b>variable</b></li><li>estimate_type</li><li>estimate</li><li>unit</li><li>variability_type</li><li>variability</li><li>interval_type</li><li>interval_lower</li><li>interval_upper</li></ul> |
| <b>cohorts</b> <ul style="list-style-type: none"><li><b>pss_id</b></li><li><b>sample_id</b></li><li><b>cohort_symbol</b></li><li>cohort_name</li></ul>   |   |   |

### S4 class performance\_metrics

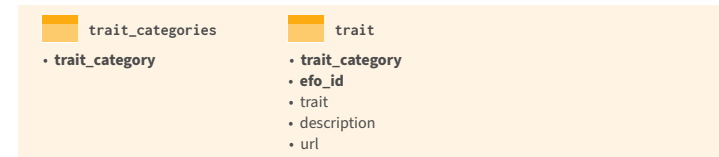
|  |  |   |
|--|--|---|
| <b>performance_metrics</b> <ul style="list-style-type: none"><li><b>pss_id</b></li><li>pgs_name</li><li>scoring_file</li><li>matches_publication</li><li>reported_trait</li><li>trait_additional_description</li><li>pgs_method_name</li><li>pgs_method_params</li><li>n_variants</li><li>n_variants_interactions</li><li>assembly</li><li>license</li><li>beta_unit</li></ul>                                 | <b>sample_sets</b> <ul style="list-style-type: none"><li><b>ppm_id</b></li><li><b>pss_id</b></li><li><b>samples</b></li><li><b>ppm_id</b></li><li><b>pss_id</b></li><li><b>sample_id</b></li><li>stage</li><li>sample_size</li><li>sample_cases</li><li>sample_controls</li><li>sample_percent_male</li><li>phenotype_description</li><li>ancestry</li><li>ancestry_description</li><li>ancestry_country</li><li>ancestry_additional_description</li><li>study_id</li><li>pubmed_id</li><li>cohorts_additional_description</li></ul> | <b>demographics</b> <ul style="list-style-type: none"><li><b>ppm_id</b></li><li><b>pss_id</b></li><li><b>sample_id</b></li><li><b>variable</b></li><li>estimate_type</li><li>estimate</li><li>unit</li><li>variability_type</li><li>variability</li><li>interval_type</li><li>interval_lower</li><li>interval_upper</li></ul> |
| <b>publications</b> <ul style="list-style-type: none"><li><b>ppm_id</b></li><li><b>pgp_id</b></li><li>pubmed_id</li><li>publication_date</li><li>publication</li><li>title</li><li>author_fullname</li><li>doi</li></ul>   | <b>cohorts</b> <ul style="list-style-type: none"><li><b>ppm_id</b></li><li><b>pss_id</b></li><li><b>sample_id</b></li><li><b>cohort_symbol</b></li><li>cohort_name</li></ul>   |   |
| 3x <b>pgs</b> {effect_sizes, classification_metrics, other_metrics} <ul style="list-style-type: none"><li><b>ppm_id</b></li><li><b>{effect_size_id, classification_metrics_id, other_metrics_id}</b></li><li>estimate_type_long</li><li>estimate_type</li><li>estimate</li><li>unit</li><li>variability_type</li><li>variability</li><li>interval_type</li><li>interval_lower</li><li>interval_upper</li></ul> |  |   |



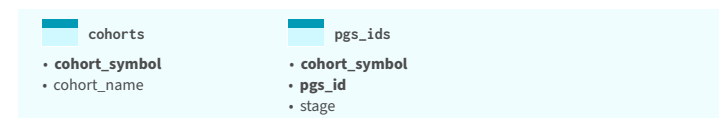
## Other S4 Entities

Besides the five PGS Catalog entities, there are three other objects that can be retrieved from the REST API: trait\_categories, cohorts and releases.

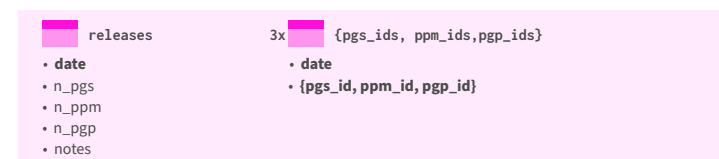
### S4 class trait\_categories



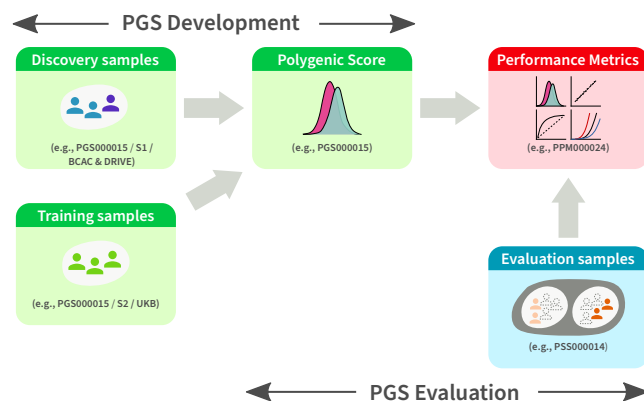
### S4 class cohorts



### S4 class releases



## PGS Construction Process



Samples and Polygenic Scores (PGS) are annotated according to their utilisation context in the PGS construction process, i.e. the stage variable in quincunx:

- "discovery"
- "training"
- "development" ("discovery" or "training")
- "evaluation"

## Cohorts, Samples and Sample Sets

### Cohorts

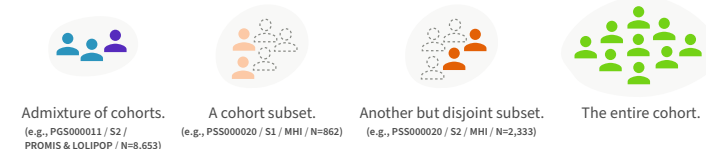
A cohort is a group of individuals with a shared characteristic. Cohorts are identified in quincunx by the cohort\_symbol variable.



### Samples

A sample is a group of participants selected from one or more cohorts. The selection from a cohort can be either a subset or its totality. Samples are not identified in PGS Catalog with a global unique identifier, but quincunx assigns a surrogate identifier (sample\_id) to allow relations between tables.

Possible compositions of samples from cohorts:



### Sample Sets

A sample set is a group of samples used in a polygenic score evaluation. Each sample set is identified in the PGS Catalog by a unique sample set identifier (PSS ID).

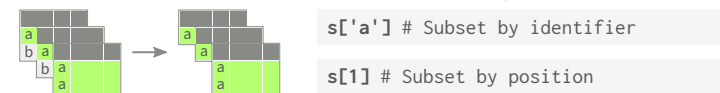


## Manipulate Cases of S4 Entities

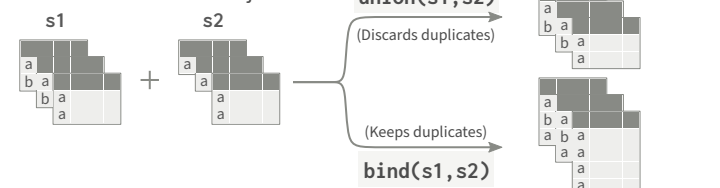
Get a scores object `s` consisting of two polygenic scores (PGS):

```
s <- get_scores(pgs_id = c('a', 'b'))
```

Subset object `s` by either identifier or position using ``[``:



Combine two scores' objects:



## Polygenic scoring file

PGS scoring files are the files provided by the PGS Catalog that allow you to compute polygenic scores for your samples. These files are hosted at the PGS Catalog FTP server: <http://ftp.ebi.ac.uk/pub/databases/spot/pgs/scores/>. They are labelled by their respective PGS Score ID (e.g. PGS000001.txt.gz). For more details please visit: <https://rmagno.eu/quincunx/articles/pgs-scoring-file.html>.

### File Format

Each scoring file contains variant identification, effect alleles and respective score weights. The file is formatted to be a gzipped tab-delimited text file, starting with a header that contains brief metadata about the score. You can read PGS scoring files into R with `read_scoring_file()`.

```
PGS000117.txt.gz
1 ## PGS CATALOG SCORING FILE - see www.pgscatalog.org/downloads/#dl_ftp_for...
2 ## POLYGENIC SCORE (PGS) INFORMATION
3 # PGS ID = PGS000117
4 # Reported Trait = Cardiovascular Disease
5 # Original Genome Build = GRCh37
6 # Number of Variants = 267863
7 ## SOURCE INFORMATION
8 # PGP ID = PGP000054
9 # Citation = Elliott J et al. JAMA (2020). doi:10.1001/jama.2019.22241
10 rsID chr_name chr_position effect_allele reference_allele effect_weight
11 rs11240779 1 808631 A G 0.00077622
12 rs1921 1 949608 A G -0.00583829
13 rs2710890 1 958905 G A -0.00182583
14 rs4970349 1 967658 T C -0.001855691
...
```

### Columns

The following table lists all possible columns in a PGS scoring file. Most columns are optional (O), a few are required (R); either the rsID or the pair chr\_name and chr\_position are required, so if one of the combinations is present the other is optional (R/O).

| Column (Requirement)       | Description                    | Example               |
|----------------------------|--------------------------------|-----------------------|
| rsID (R/O)                 | dbSNP Accession ID             | "rs554219"            |
| chr_name (R/O)             | Chromosome name                | "11"                  |
| chr_position (R/O)         | Chromosome position            | 69516874              |
| effect_allele (R)          | Effect allele                  | "G"                   |
| reference_allele (O)       | Reference allele               | "C"                   |
| effect_weight (R)          | Variant weight                 | 0.117                 |
| locus_name (O)             | Locus name                     | "CCND1"               |
| weight_type (O)            | Type of weight                 | "log(OR)", "beta_cox" |
| allelefrequency_effect (O) | Effect allele frequency        | 0.410                 |
| is_interaction (O)         | Variant interaction?           | TRUE or FALSE         |
| is_recessive (O)           | Recessive inheritance model?   | TRUE or FALSE         |
| is_haplotype (O)           | Is effect allele a haplotype?  | TRUE or FALSE         |
| is_diploptype (O)          | Is effect allele a diploptype? | TRUE or FALSE         |
| imputation_method (O)      | Imputation method              | TODO                  |
| variant_description (O)    | Variant description            | TODO                  |
| inclusion_criteria (O)     | Score inclusion criteria       | TODO                  |
| OR (O)                     | Odds Ratio                     | 1.12                  |
| HR (O)                     | Hazard Ratio                   | 1.08                  |