INTERVENE

D2.2 Plan of data standards for genetic risk scores

Deliverable

Dissemination level: Public

Part of:

WP2:

Data harmonisation, integration, and evaluation of genetic scores

Project summary	
Project acronym:	INTERVENE
Project full title:	International consortium for integrative genomics prediction
Project Coordinator	Institute for Molecular Medicine Finland FIMM, University of Helsinki; Prof. Samuli Ripatti and Dr. Andrea Ganna
Project start date:	1.1.2021
Project end date:	31.12.2025
Project duration:	60 months
Action type:	RIA
Call identifier:	H2020-SC1-FA-DTS-2018-2020 (Trusted digital solutions and Cybersecurity in Health and Care)
Grant number	101016775
Document descriptors	
Deliverable No.	2.2
Work package	WP2
Deliverable lead	UCAM
Contributors	UCAM, EMBL-EBI, UTARTU, FIMM
Dissemination level	PUBLIC
Expected delivery date	30/06/2021
Submission date	30/06/2021

Change history log

Version	Changes made	Date	Prepared by	PC approved
0.1	**Outline created**	**21/04/2021**	**EMBL-EBI**	
0.2	**Draft written**	**07/06/2021**	**EMBL-EBI, UTARTU**	
0.3	**Draft updated**	**11/06/2021**	**EMBL-EBI, UTARTU, UCAM**	
1.0	**Minor formatting changes**	**22/06/2021**	**UH-FIMM**	**Yes**

Contents

1. Introduction

This deliverable relates to Task 2.2 Standards for genetic risk scores. It addresses WP2 Objective 1 and WP4 Objective 1 and is crucial for the interoperability of the polygenic scores and analysis platforms used and developed by the project. The aim of this task is to create a standardized format for risk scores and their meta-data, as well as standardized methods for reporting risk score predictive ability.

Beyond the INTERVENE project, alignment with the broader genetic risk score field will facilitate interoperability with the community at large. Therefore, we have decided to base our plan for standards on the recently published Polygenic Risk Score Reporting Standards (PRS-RS; Wand et al., 2021)[1]. The PRS-RS is a framework which describes the minimal information that is required to interpret and evaluate genetic risk scores. It has been developed as a collaboration by the INTERVENE partners at the Polygenic Score Catalog (UCAM, EMBL-EBI) and the ClinGen Complex Disease working group. We expect that INTERVENE's use of this internationally supported standard will inform the practical aspects of implementation at scale and across international boundaries.

To unify reporting of predictive ability for genetic risk scores, standardized methods and formats are introduced. The minimal format of the scoring file itself is based on the existing format used by the Polygenic Score Catalog (https://www.pgscatalog.org/), itself a partner of INTERVENE where it will contribute and share novel scores.

Task 2.2 also includes the development of a tool for the calculation of scores on genotype data stored in indexed VCF v4.2 file format. VCF is the format specified in Deliverable 2.1 (Data resources and harmonisation) as the standard format to be used for genotype data. Currently, tools are available to calculate risk scores using plink format (bfile or pgen) as well as bgen format files as input. In this deliverable, we provide a tool for rapid risk score calculation using vcf format as an input file.

Objectives addressed:

- WP2 Objective 1 - To harmonize genetic and other -omics data with a specific focus on standards for generation and reporting of genetic scores
- WP4 Objective 1 - To expand the Polygenic Score Catalog to include genetic scores arising from INTERVENE and their relative performance on INTERVENE-participating studies.

2. Methodology

To ensure maximal interoperability and compliance with established community standards we have elected to base the standards used in INTERVENE on existing state of the art standards, the PRS-RS and the PGS Catalog scoring file format for data representation. While these may change over time, in part due to the knowledge gained from the INTERVENE project, they represent the state of the art in the field. The currently published PRS-RS (Wand et al, 2021) is the minimal reporting information required;

[1]Wand et al., 2021; https://www.nature.com/articles/s41586-021-03243-6?proof=t

however, INTERVENE will continue to monitor the state of the art in the field, updating the PRS-RS as necessary once new requirements arise (see below).

Data for which standards are required:

1. Score development and predictive ability
2. Polygenic Scores (linear combination of variant alleles and weights)

3. Results

3.1 Plan of standards for genetic risk scores

The PRS-RS is a set of standard items specifying the minimal criteria that need to be described to accurately interpret a genetic risk score and reproduce results throughout the score development process.

Appendix 1 contains the contents of Table 1 from Wand et al. all of which are directly relevant to INTERVENE's cohorts. Reporting items span detailed descriptions of study populations, statistical methods for the development and validation of PRSs and considerations for the potential limitations of these scores. Items are organised into key components along the developmental pipeline of genetic risk scores for clear interpretation and to encourage their documentation from the inception of the study.

Scores developed and evaluated by INTERVENE partners should document/record the meta-data items as outlined in Appendix 1. Specifically, to facilitate WP4 Objective 1 (to expand the Polygenic Score Catalog to include genetic scores arising from INTERVENE and their relative performance on INTERVENE-participating studies) the PRS-RS data items should be captured in the structured fields used by the PGS Catalog (described in Supplemental Table 2 of Wand et al, Appendix 2).

Structuring information according to these fields will facilitate future population of the PGS Catalog curation template, and upload into the PGS Catalog thereby disseminating them and demonstrating publicly the relevance of the standards and Intervene adherence to it and future developments needed. In line with the PGS Catalog, we propose ancestry information should be captured using the standardised framework developed by the NHGRI-EBI GWAS Catalog[2]. These items are further described on the PGS Catalog website (https://www.pgscatalog.org/about/#submission).

Genetic Risk Score

The genetic risk score itself is described in a 'scoring file', which in order to be usable must contain at minimum a unique identifier for the variants and specify an effect allele to which an effect size is ascribed. To further ensure interoperability we propose to use the standard fields and information content used by the PGS Catalog, which in turn were developed to closely resemble existing file formats used to calculate scores in common software packages (e.g. PLINK).

In addition to minimal required fields (chromosomal position; allele and effect weight), the reference genome assembly version (GRCh37/38) must be specified in the file header. Each scoring file should be a

[2] Morales et al., 2018; https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1396-2

tab-delimited text file. We propose four minor modifications to the file format used by the PGS Catalog (https://www.pgscatalog.org/downloads/)[3] to facilitate greater clarity and utility of scoring files.

1. The PGS Catalog uses the term 'reference allele' as the column header for the non-effect allele, we propose renaming this with the term 'other allele' to disambiguate the required field content (the term 'reference allele' may be used to refer to the allele in the reference genome, which may or may not be the non-effect/other allele). This terminology is also consistent with the GWAS summary statistics standard proposed by the NHGRI-EBI GWAS Catalog[4]. This disambiguation will facilitate ease of communication and avoid potential confusion.

2. The PGS Catalog has designated the 'other allele' field as 'optional, but highly recommended', primarily because it relies on author submission of previously generated scores where the other allele may not have been recorded. INTERVENE has the opportunity to mandate inclusion of this information when scores are generated, which will facilitate greater utility of the scoring files (for example the vcf calculation tool presented in this deliverable requires this field).

3. The PGS Catalog specifies that either rsID or chromosomal position (chr_name, chr_position) are used as the variant identifier. For the same reasons as above (2) and to reduce external mapping dependencies, INTERVENE mandates inclusion of chromosomal position information (chr_name and chr_position) for all variants.

4. The file header is modified to remove items that are unique to the PGS Catalog e.g. PGS Catalog specific identifiers.

These required fields and additional recommended fields are described in Table 1a (scoring file header) and Table 1b (score file column headers). We will monitor for possible future changes to this file format based on our experience in the project. For example, we also consider adding the phenotype definitions which are used for PRS testing or a direct link to that information to simplify reproducibility.

[3] Lambert et al., 2021; https://www.nature.com/articles/s41588-021-00783-5
[4] Buniello et al., 2019; https://doi.org/10.1093/nar/gky1120

Table 1a. Scoring file header

Proposed:	**Modified from:**
INTERVENE Scoring File header	**PGS Scoring File header**
### INTERVENE SCORING FILE	### PGS CATALOG SCORING FILE - see www.pgscatalog.org/downloads/#dl_ftp for additional information
### GENETIC RISK SCORE INFORMATION	## POLYGENIC SCORE (PGS) INFORMATION
	# **PGS ID** = *PGS idenfier, e.g. 'PGS000001'*
# **PGS Name** = *PGS name, e.g. 'PRS77_BC'*	# **PGS Name** = *PGS name, e.g. 'PRS77_BC' - optional*
# **Reported Trait** = *trait, e.g. 'Breast Cancer'*	# **Reported Trait** = *trait, e.g. 'Breast Cancer'*
# **Original Genome Build** = *Genome build/assembly, e.g. 'GRCh38'*	# **Original Genome Build** = *Genome build/assembly, e.g. 'GRCh38'*
# **Number of Variants** = *Number of variants listed in the PGS*	# **Number of Variants** = *Number of variants listed in the PGS*
## SOURCE INFORMATION	## SOURCE INFORMATION
# **Author information** = *Information about the creator(s) of this score*	# **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*

Table 1b. Scoring file column headers. *indicates changes to the format used by the PGS Catalog. 'Other allele' is referred to in the PGS Catalog scoring file as 'reference allele', referred to here as 'other allele' for clarity (to disambiguate from reference genome allele). The other allele field is 'optional but strongly recommended' by the PGS Catalog, we propose mandating inclusion of this information. The PGS Catalog mandates either (rsID) or (chr_name and chr_position) as the variant id, we are specifying that chr_name and chr_position are mandatory.

Column Header	Field Name	Field Description	Field Requirement
rsID	dbSNP Accession ID (rsID)	The SNP's rs ID	Optional, but recommended*
chr_name	Location - Chromosome	Chromosome name/number associated with the variant	Required*
chr_position	Location - Position within the Chromosome	Chromosomal position associated with the variant	Required*
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'.	Required
other_allele*	Other Allele*	The other allele(s) at the loci ie the non-effect allele	Required*
effect_weight	Variant Weight	Value of the effect that is multiplied by the dosage of the effect allele ('effect_allele') when calculating the score.	Required
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 something like an **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.	
is_diplotype			Optional
imputation_method	Imputation Method	This described 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 gets included into the PGS (e.g. if it depends on the results from other variants).	Optional
Extra columns:			
OR/HR	Odds Ratio [OR], Hazard Ratio [HR]	Author-reported effect sizes can be supplied to the Catalog. If no other *effect_weight* is given the weight is calculated using the log(OR) or log(HR).	Optional
allelefrequency_effect_Ancestry	Population-specific effect allele frequency	Reported effect allele frequency in a specific population (described by the authors).	Optional

3.2 Risk score calculator for vcf formats

Prof. Reedik Mägi has developed a risk score calculator named faSt Tool for gEnetic Risk scOre calculatIon using biobank Data (in short, STERIOD, https://genomics.ut.ee/en/tools/steroid). It requires vcf format as input for generating scores and it can use both best guess as well as maximum posterior probability data of imputed genotypes. As an input file, LDpred format (https://github.com/bvilhjal/ldpred) is required (Table 2). More information regarding STEROID, its command line options and examples can be found on tool's home page.

Table 2. Input file format required for STEROID risk score calculator. nt1 is the non-effect (other) allele and nt2 is the effect allele. The variant ID header 'sid' is not a typo. Raw_beta is the original beta from GWAS study and ldpred_beta a posterior mean from ldpred gibbs sampling, but any beta can be given in the last column for genetic risk score calculation.

chrom	pos	sid	nt1	nt2	raw_beta	ldpred_beta
chrom_1	752566	rs3094315	G	A	6.00E-03	-4.25E-05
chrom_1	768448	rs12562034	A	G	2.70E-03	-1.39E-05
chrom_1	779322	rs4040617	G	A	6.20E-03	-4.66E-05
chrom_1	785989	rs2980300	T	C	5.50E-03	-3.47E-05

Conclusions

We have evaluated the state of the art standard PGS-RS and it is broadly appropriate for INTERVENE's use cases and datasets as we understand them at present. We are aware of cases where the standard should be modified and have provided our reasoning as to why these modifications are necessary (Section 2a, Table 1a, 1b). As INTERVENE progresses we will review the standard's use for INTERVENE and report any additional or changed needs for the standard with the aim of improving the standard via our experience of real world use.

Appendix 1: Table 1 from Wand et al.

Table 1 Polygenic Risk Score Reporting Standards (PRS-RS)

From: Improving reporting standards for polygenic scores in risk prediction studies

Reporting standard		Description
Background	**Study type**	Specify whether the study aims to develop and/or validate a PRS. When externally validating or combining previously published PRSs or integrated risk models, include identifier(s) of original PRS (PMID, PGS Catalog ID).
	Risk model purpose and predicted outcome	Specify what the risk model is intended to predict and the purpose. This includes intended use (risk prediction, diagnostic, prognostic, or therapeutic modalities), predicted outcome (if a clinical feature or endpoint within a specific disease) and the current models available for that outcome.
Study population and data Many risk score studies involve multiple populations and cohorts that can be used in different stages of PRS and risk score development and evaluation. Each of the populations used (for example, training, validation and subgroup analysis samples) in the manuscript should be defined using this common set of descriptors.	**Study design and recruitment**	For each of the datasets describe the study design (for example, cohort, case–control, cross-sectional), eligibility criteria, recruitment period and setting (for example, method and years) and follow-up. State whether the data are primary or secondary data. If secondary analysis, include the full reference to the original study.
	Participant demographics and clinical characteristics	Include the distribution of demographic information in each dataset (and the combined total if relevant) used to generate a single risk model (whether a single sample set, or the summary of combined samples) including the mean, standard deviation and range. This should at minimum include age, sex and any other characteristics relevant to describe the study population or the performance of the model. Provide demographics stratified by case–control status, if applicable.
	Ancestry	Include the ancestral background distribution of each sample population used during PRS development and validation (including those from any GWAS summary statistics that were included), and the data source of this ancestry information (for example, self-report, genotyping). Ancestry information should be reported using the standardized framework developed by the NHGRI-EBI GWAS Catalog[1] with detailed information beyond this when available. When combining samples from multiple studies, aggregate ancestral distribution information is sufficient. The method of ancestry inference should be provided.
	Genetic data	Provide the method for acquiring genetic information (for example, sequencing, genotyping) in each sample, including information about genome build and technical assay details. If imputed, specify the imputation panel and give ancestry information. Report any relevant quality control, including imputation quality filters to exclude low-quality imputed SNPs. If parameters were selected from another study, include reference (PMID, GWAS Catalog ID).
	Non-genetic variables	Define any non-genetic variables that were included in the risk model, provide variable definitions and measurement (for example, assay, ICD codes, e-phenotyping algorithms, chart review, self-report). Indicate the scale of each variable, for example, dichotomous, continuous, categorical or ordinal. Explicitly state which variables are included in the final model.
	Outcome of interest	Define the predicted outcome(s) of interest and report distribution. If the predicted outcome is a clinical feature or end-point within a specific disease, provide the criteria used to define that disease membership. Include details on how information was ascertained (for example, ICD codes, e-phenotyping algorithms, chart review, self-report). Transformation of continuous data into binary, ordinal, or categorical outcomes should be detailed with justification. State whether the predicted phenotype of the polygenic score is the same as or different from the predicted outcome of the risk model. Provide justification for differences, if applicable.
	Missing data	State explicitly how missing data were handled for all variables included in the model. If imputation was used, include detailed of the approach used and any subsequent filtering or post-processing.
Risk model development and application Describe methods used to form the final PRS and/or risk model. Samples in this stage of the analysis should be denoted 'Score development' or 'Training', and described in 'Study population and data'.	**PRS construction and estimation**	Describe how genetic data were included in the PRS. Authors should detail criteria used to determine inclusion in the model for all variants. Define how the variants were selected, weighted and combined into a single score. If the PRS was derived from another study include the reference (PMID, PGS Catalog Score ID).
	Risk model type	Detail statistical methods used to estimate risk, either relative or absolute, from the continuous risk score distribution. Detail whether risk is cumulative or cross-sectional, with appropriate comparison groups if relative risk is presented. Report time until predicted risk (for example, 5 years, 10 years, lifetime). In an absolute risk model, state the time until the predicted event and the prevalence or incidence of the predicted outcome in the general population.
	Integrated risk model(s) description and fitting	State the procedure used to develop the risk models that includes non-genetic and/or genetic variables other than the PRS. If the model(s) was selected for optimal performance, describe measures used to assess performance. Explicitly state all variables used in each risk model.

Limitations and clinical implications Discuss the broader context of the study and risk model.	**Risk model interpretation**	Summarize the risk models in terms of what they predict, how well and in whom. Explicitly mention the incremental performance of the PRS and/or combined risk model in comparison to conventional risk models, as well as the performance of the PRS and risk model alone. Conventional risk models might include demographic (age, sex), disease-specific risk factors and/or family history of disease.
	Limitations	Outline limitations of the study with relevance to the results, discuss the effects of these limitations on the interpretation of the risk model and any downstream replication efforts needed. Common considerations include: study design restrictions, use of a surrogate outcome, ascertainment biases, the distribution of participant-level traits (ancestry, age, comorbidities), accuracy or specificity of outcome data, and any statistical considerations. Note and discuss the effects of any unknown reporting items from previous sections.
	Generalizability	Discuss the intended target groups or populations this score may be applied to and explicitly address any issues with generalizability beyond the included populations. Discuss whether the study externally validates the score and/or model, or if the sample is limited with respect to ancestry, age or other variables.
	Risk model intended uses	Discuss whether there is an intended clinical use or utility to the risk model. If so, discuss the 'clinic readiness' and next steps with respect to the interpretation, limitations and generalizability of the model. Discuss how the predictive ability of the model compares with current standards of care or other published work (such as existing PRSs) on predicting the outcome of interest.
Data transparency and availability		Information sufficient to calculate the PRS and the risk model(s) on external samples should be made freely available. For genetic variables this would include information about the variants (for example, rsID, chromosomal location, effect allele and the effect weight) that comprise the score; PRSs with this information should be deposited in the PGS Catalog for findability and to promote reuse and comparison with other established scores. Weights for non-genetic variables should also be provided to make the risk model calculable.

Appendix 2: Supplemental Table 2 from Wand et al

PRS-RS		Mapped PGS Catalog data items			Comments on mapping
Section	**Reporting Item**	**Data Type**	**Field**	**Description**	
Risk Score Background	**Risk Model Purpose & Predicted Outcome**	**Score**	Reported Trait	Phenotype the PGS predicts	*Terminology is adapted to not be limited to clinical traits. PGS Catalog curators also map the trait terms to an ontology to organize scores and consistently define phenotypes.*
			Mapped Trait(s)	Ontology term/mapping of the Reported Trait using the Experimental Factor Ontology (EBI)	*Mapped by PGS Catalog curators to ensure consistency*
Study Populations	**Study Design & Recruitment**	**Sample** *Samples in the PGS Catalog are described consistently and can be tagged as being used in one of the three main stages of a PGS/PRS study: samples used to identify variant associations (GWAS), Score Development, or Testing. Samples in the Catalog are also be annotated with the Cohort that they originate from to prevent future applications to training samples (overfitting).*	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).	
			Additional Sample/Cohort Information	relevant information not captured by the structured fields	*This information isn't required or structured in the PGS Catalog, but can be recorded as free text*
	Ancestry		Broad Ancestral Category	Author reported ancestry is mapped to the best matching ancestry category/categories from the GWAS Catalog framework (Table 1 of Morales et al. (2018)).	*Recorded using the GWAS Catalog's framework from Morales et. al (2018)*
			Ancestry	A more detailed description of sample ancestry that usually matches the most specific description provided 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).	
	Participant Demographics		Age of Study Participants	A summary (mean/median, range/confidence intervals) of study participants ages.	*The PGS Catalog captures age and sex in defined fields, as they are the most commonly reported descriptions of participants.*
			Percent of participants who are Male	Author reported percentage of participants in the described sample that are male.	
	Non-Genetic Variables		Detailed Phenotype Descriptions	A description of how the phenotype was measured or defined (e.g. ICD codes used to identify cases/phenotypes in EHR data). *(This can include information about the non-genetic variables included in the risk model)*	*Non-genetic variables included in the risk model are also recorded in the* Covariates field of the **Performance Metrics** data type
	Outcome of Interest				
	Genetic Data	**Score**	Genome Build	*Original genome build the variants/PGS are associated with*	
			Score Development Details	A description of the relevant inputs and parameters relevant to the PGS development method/process.	*If the score is requires any or is specific to any genotyping methods or imputation it can be recorded here.*

Risk Score Development & Application	**Risk Model Type**	**Score**	Score Development Method	The name or description of the method or computational algorithm used to develop the PGS.	
	Polygenic Risk Score Construction & Estimation				
	Integrated Risk Model(s) Description and Fitting		Score Development Details	A description of the relevant inputs and parameters relevant to the PGS development method/process.	
Risk Score Evaluation	**Risk Model Predictive Ability**	**Performance Metrics**	PGS Effect Size (per SD of polygenic score)	Standardized 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.	*Metrics can all be annotated with a confidence interval. Performance Metrics must link to a single PGS, and set of Samples (identified by a PSS ID). Multiple scores can be compared on the same set of samples (PSS ID).*
	Risk Model Discrimination		PGS Classification Accuracy	Examples include the Area under the Receiver Operating Characteristic (AUROC) or Harrell's C-index (Concordance statistic).	
	Risk Model Calibration		Other metric(s)	Metrics that do not fit into the other two categories. Examples include: R2 (proportion of the variance explained), or reclassification metrics.	
	Subgroup Analyses	**Sample, Performance Metrics**	*Multiple*		*Each subgroup analysis can be recorded as a separate set of Sample descriptions with linked Performance Metrics*
Translation	**Risk Model Interpretation**	**Sample, Performance Metrics**	*Multiple*		*Comparisons to existing risk models can be recorded as Performance Metrics for the same sample employing different sets of covariates (example PGS evaluated alone or after adjusting for age+sex: http://www.pgscatalog.org/score/PGS000036/).*
	Generalizability				*Generalizability of each PGS/PRS can be captured in the Catalog, as performance metrics for the same score evaluated in different samples and/or ancestries is captured on a single page. Subsequent evaluations of multiple PGS can also be performed and added to the catalog as a benchmarking paper (example benchmarking paper of 4 existing scores on samples from 3 ancestry groups: http://www.pgscatalog.org/publication/PGP000083/)*
Data Transparency and Availability		**Scoring File, PGS Catalog Downloads, Terms & Licenses**	*Multiple*		*The PGS Catalog provides the variant file necessary for applying the PRS to new samples, the metadata recorded supports application and evaluation.*