# Improving Reporting Standards for Polygenic Scores in Risk Prediction Studies

ClinGen
Clinical Genome Resource

## 1 Introduction

State whether study describes risk score development or external validation of a *previously published risk score.
*Include identifiers (PMID, PGS Catalog ID)

Define the purpose of the polygenic risk score.
Is it for risk prediction? Diagnostic? Prognostic? Therapeutic?

**Why?**
Sufficient detail is needed to understand the nuance in measured risk.

E.g., there can be a mismatch in risk-score predicted outcome (e.g. CAD) and the measured phenotype of interest (e.g LDL).

Specify the predicted clinical end outcome.
Is this a feature/endpoint of a specific disease?

**Additional Recommended Practices:**
Specify the goal in predicting this outcome, and what constitutes a "good" predictor. E.g., present AUC from other clinical risk models.

## 2 Methods: Study Design

Include study design details.

- **Study Type** — Cohort study? Case-control? Cross-sectional?
- **Clinical End Outcome** — Is this defined by incidence or prevalence?
- **Data Source** — Are these primary or secondary data?
- For secondary analyses, include a reference to the original study/studies.

State recruitment details, i.e. method & years.
If performing secondary analyses, reference recruitment details.

**Why?**
The "who, where, and when," of risk depend on the population used to derive the risk model.

Include details defining the phenotype of interest.

- **How was the phenotype defined?** — Provide inclusion & exclusion criteria for the predicted outcome. If the predicted outcome is a feature or endpoint of a specific disease, provide criteria to define that disease. State the number of cases and controls.
- **How was information ascertained?** — ICD codes? Phenotyping algorithms? Chart review? Self report? Transformation of continuous data into binary outcomes should be detailed for reproducibility.

**Additional Recommended Practices:**
- Performance should not be investigated in case-control studies.
- Explain when prediction of prevalence is justified/clinically relevant.
- For validation studies, provide justification if the predicted outcome differs from phenotype of interest in score development.

## 3 Methods: Variable Definitions

Ancestry: Include the distribution* of ancestral background and the data source. Use the NHGRI-EBI standardized reporting framework.

Sex: Include counts and percentages.
State if sex was inferred from self-report or genetic information.

Age: Include age distribution using the mean, standard deviation and range.
*Include by case/control status, if applicable.

- **Missing Data** — How was this handled?
- **Risk Model Parameter Specifications** — State all terms used in the final risk model.
- **Risk Model Genetic Data Acquisition** — Include data source. (Sequencing vs. genotyping?) Include genome build and details on imputation.
- **Clinical Variable Definition(s)** — Provide inclusion & exclusion criteria for any non-genetic variables in the risk model. Include data source. State if dichotomous or continuous.
- **Sub-Analyses** — Provide inclusion & exclusion criteria used to stratify or subset the sample. Include data source and justification for any cut-offs. State the number of cases included in sub-analyses.

ClinGen
Clinical Genome Resource

**Why?**
The details needed to reproduce or critically analyze a risk score are often underreported.

## 4 Methods: Analysis

- **PRS Estimation** — Describe how data are transformed into a score.
- **Model Fitting Procedure** — Describe how the final model was selected.
- **Model Type** — Describe methods used to estimate risk.
- **Calibration** — Describe measures used to assess calibration.
- **Discrimination** — Describe measures used to assess discrimination.
- **Validation** — If performed, outline procedures for validation.
- **Sub-Group Analyses** — If performed, provide details.

## 5 Results

Risk Score Distribution
Include a general description of the risk score, as well as model fit measures.

Risk Score Predictive Ability
State if risk model output is in terms of absolute or relative risk. Provide enough detail for readers to compute measures of predictive ability.

**Additional Recommended Practices:**
Provide the risk score distribution in a visual display, showing overlap in the distributions for cases and controls. Explicitly state the summary statistics the algorithm produces (e.g. hazard/odds ratio), and any reference levels used.

Risk Score Discrimination and Calibration
Include metrics assessing discrimination & calibration, and whether other variables are included beyond the risk score.

Subgroup Analyses
Report all measures for any subgroup analyses.

Risk Score Validation
*Report all measures conducted in a validation set.

## 6 Discussion

Risk Model Interpretation:
What does it predict? How well? In whom?

- Predicted outcome should be consistent with the introduction.
- Explicitly mention the risk model's performance beyond non-genetic factors.
- Target population should reflect the sample used in model development.

Limitations:
Outline limitations and their impact to interpretation of the risk score.

- Study Design Restrictions
- Ascertainment Biases
- Distribution of participant level traits
- Accuracy / specificity of phenotyping
- Any unknown reporting items

**Why?**
By explicitly describing the risk model's interpretation and outlining potential limitations to the generalizability of their model, authors will empower readers and the wider community to better understand the risk score and its relative merits.

Generalizability:
Discuss which populations this score may be applied to, and any issues with generalizability.

Risk Score Intended Purpose:
If there is an intended clinical purpose, discuss "clinic readiness." How is the model benchmarked against current standard of care?

**Additional Recommended Practices:**
-Interpretation: Discuss whether model(s) were adjusted for family history. Give AUC for all models tested, including one based on age and sex only.
-Limitations: Explicitly state any overlap in samples between GWAS, testing, and validation datasets. Ideally there should be none.
-Generalizability: Discuss whether the score replicates previous findings or has been externally validated.
-Data Transparency & Availability: Information can be published in the PGS Catalog (pgscatalog.org) to promote re-use.

Data Transparency & Availability:
Information sufficient to calculate the PRS and/or risk model should be made available.

Read the full paper at:
https://tinyurl.com/y7fwz64e
clinicalgenome.org | pgscatalog.org
Updated June 15, 2020

ClinGen
Clinical Genome Resource