Poster: Private information leakage from polygenic risk scores

Kirill Nikitin
Columbia University & New York Genome Center
kirill.nikitin@columbia.edu

Gamze Gürsoy
Columbia University & New York Genome Center
gamze.gursoy@columbia.edu

***Abstract*—Polygenic Risk Scores (PRSs) estimate the likelihood of a person to develop diseases based on their genetic variations. Individual PRS values are frequently shared with results of clinical studies or on online health platforms. We demonstrate how a profit-seeking actor, such as an insurance provider, can exploit PRSs to recover the associated genotypes of individuals and to either de-anonymize these individuals or infer their health traits. By framing genotype recovery as the subset-sum problem with side information from population statistics, we show how to reconstruct a significant portion of an individual's genome from their individual PRS values with 95% accuracy. The predicted genotypes or the PRS itself are then sufficient to identify the individual or their relatives in genealogy databases or public anonymized biobanks.**

1. Introduction

A polygenic risk score (PRS) estimates an individual's genetic susceptibility to complex traits. It is increasingly used in clinical settings to inform risk stratification, early intervention, and personalized care. Leading genetic-testing companies, such as 23andMe and Nebula Genomics, already provide PRS estimates to consumers in their genetic reports.

While ethical and policy debates on the commercial use of PRSs are emerging, the privacy risks of sharing individual scores remain largely unexamined. A PRS is summary data by nature, as it is a single number that represents the cumulative effect of multiple genetic variants. It is not uncommon for genetic studies to release anonymized individual scores for reproducibility purposes or for customers of genetic-testing companies to post their scores online to seek health advice. However, even summarized or obfuscated genetic data can be highly sensitive and can lead to re-identification of patients [1] or exposure of their health status [2]. This raises the question of whether PRSs expose private genetic information, under what conditions they do so, and in what scenarios this leakage can pose privacy risks. We answer these questions in our study.

2. Background and threat scenarios

The human genome is organized into chromosomes, each of which comes in paris—one inherited from the mother and one from the father. Variations in these paired DNA sequences contribute to individual differences in physical traits and disease susceptibility. The most common form of genetic variation is the single nucleotide polymorphism (SNP), a substitution of a single nucleotide at a specific position in the genome relative to a standardized reference sequence. At a given SNP site, if both copies of a person's genome match the reference, the individual is assigned a genotype of "0." If one chromosome differs from the reference whereas the other matches, the genotype is "1." If both differ from the reference, the genotype is "2." Genome-wide association studies (GWAS) examine SNPs across the genomes of large populations to identify statistical associations between genetic variants and specific diseases.

A PRS model is derived from the significant associations found in GWAS. It is calculated as a linear combination of the genotypes $g \in \{0, 1, 2\}$ of the associated SNPs and effect weights β. The effect weights are real numbers with varying degree of precision (the number of decimal places), which are released alongside other model metadata. The score is normalized by the number of SNPs N and ploidy P:

$$PRS = \frac{\sum_{i=1}^{N} \beta_i g_i}{P \cdot N} \tag{1}$$

In our threat model, an attacker accesses a panel of PRSs with the corresponding metadata of anonymous or known individuals, which might be released by a research study or posted online. The attacker attempts to infer the associated genotypes of these individuals and to either de-anonymize them or uncover their sensitive phenotypic information.

3. Recovering genotypes from PRS

In order to infer an individual's associated genotypes, the adversary needs to find which effect weights β_i sum up to a target score. This task is a variation of the classic subset-sum problem, which, despite its NP-hardness, can be efficiently solved under certain conditions. We reduce the genotype-recovery task to an instance of the subset-sum problem by defining the effect weights as the number set and the PRS value as the target sum. The number of times each effect weight (0, 1, or 2) is used in the solution determines the genotype for each corresponding SNP.

The hardness of the subset-sum problem is traditionally defined by using the concept of density [3]. It represents the ratio between the size of the number set and the length of

the bit representation of the largest weight. Higher density indicates that there are more weights relative to the bit length of the largest weight and, hence, a tighter distribution of possible subset sums. Problems with a higher density are more difficult to solve because multiple subsets can yield the same sum. We adapt this concept to genetic data and define the density d of a PRS instance as $d = \frac{N}{\log_3(max_{1 \leq i \leq N} \texttt{ decimal}(\beta_i))}$, which enables us to assess what problem instances are tractable.

We develop a dynamic programming algorithm with the meet-in-the-middle optimization and additional heuristics [4] to infer the SNP genotypes from a PRS. This algorithm, however, identifies *all* valid solutions, i.e., all possible genotype sets that result in the target PRS. To assess the plausibility of each potential solution, we calculate a log-likelihood score by using genotypes frequencies from the target individual's population. We compute the sum of the log-probabilities for the identified genotypes in each solution and select the solution with the highest total sum. The core idea is that a solution that closely aligns with the population average is more likely to be correct. Finally, we incorporate continuous log-likelihood estimation directly into the dynamic-programming algorithm. For each intermediate sum, we store pointers only to the subsets that result in the highest likelihood, thereby reducing memory complexity from exponential to pseudo-polynomial.

PRS models frequently share overlapping SNPs, as a genetic mutation can affect multiple traits. When an adversary gets access to multiple PRS values of an individual, e.g., from a genetic report [5], they can exploit this overlap to improve recovery accuracy. The strategy is to first predict genotypes for the PRSs with fewer SNPs and to retain the genotypes of the overlapping SNPs from the previous solution for each subsequent PRS. When incorporating solutions from a previous PRS, if the solution for the current PRS fails, it suggests that the integrated genotypes might be incorrect, which we can revisit and correct.

4. De-anonymization and phenotype inference

PRSs are designed to stratify individuals by their genetic risk. Besides clinical utility, this can make PRSs a powerful identification tool. An attacker can re-identify anonymous individuals or their close relatives, e.g., study participants or online forum users, by uploading their recovered genotypes to a genealogy database and finding genetic matches. Genealogy databases do not provide direct access to individuals' genomes but they allow users to query the database.

Moreover, if the attacker has access to an anonymized genotype-phenotype database, e.g., UK Biobank, and a PRS value and model of a known individual for some trait, the attacker can link the PRS to a database sample without computationally intensive genotype recovery and learn other sensitive traits that the individual might have [6]. The linking is simple: the attacker calculates the PRS for every sample in the database and finds the one that matches the target. This approach relies on determining how unique PRS values are across individuals which we evaluate in our experiments.

5. Evaluation

Solvability. We first analyzed all PRS models published in the PGS Catalog [7] to assess the proportion that would be vulnerable if an individual's PRS were shared. Following prior work on the subset-sum problem, we determined PRS instances with density $d < 2.5$ to be solvable. Based on this density, we found that 454 out of 4,723 published PRS models were vulnerable to genotype recovery. The largest vulnerable PRS model included 95 SNPs and had the effect weights with up to 21 decimal digits, whereas the median weight precision across all models was 15 decimal places.

Genotype recovery. We evaluated genotype recovery by using whole-genome sequencing data from the 1000 Genomes Project (2,535 samples) and a panel of 298 PRSs, up to 50 SNPs each, from the PGS catalog for various diseases. The total number of encompassed SNPs was 4,821 SNPs, of which 2,654 were unique. For each sample, we first calculated the PRS values and then attempted to predict the original SNPs genotypes by using our algorithms. We achieved a median genotype prediction accuracy of 94.6% with 2,600 SNPs predicted on average. The baseline of predicting the most common genotypes in the population for each SNP achieved only 70% accuracy. We also observed that genotype prediction was the least accurate for the SNPs with equally likely genotypes in the population.

De-anonymization and score uniqueness. We used the KING-robust algorithm [8] and the 1000 Genomes dataset to emulate genealogy search. Despite only 2,600 partially correct predicted SNPs, we were able to link each individual to themselves with 100% accuracy. We were also able to link individuals to their first- and second-degree relatives (68 samples) with >80% precision and recall. Finally, we studied the UK Biobank dataset (500K samples) and found that a single PRS based on 27 SNPs, on average across all weight precisions, sufficed to uniquely identify 95% of individuals. Even when the scores overlapped, the same score was often shared by only few individuals. In PRS models with 14 SNPs, the median anonymity-set size was two.

References

[1] R. Wang *et al.*, "Learning your identity and disease from research papers: information leaks in genome wide association study," in *ACM CCS*, 2009.

[2] D. R. Nyholt *et al.*, "On Jim Watson's APOE status: genetic information is hard to hide," *European Journal of Human Genetics*, 2009.

[3] J. C. Lagarias and A. M. Odlyzko, "Solving low-density subset sum problems," *Journal of the ACM*, 1985.

[4] E. Horowitz and S. Sahni, "Computing partitions with applications to the knapsack problem," *Journal of the ACM*, 1974.

[5] "23andMe launches new genetic reports on common forms of cancer," Online, 2024, acc: 2025-04-13.

[6] The Guardian, "Private UK health data donated for medical research shared with insurance companies," Online, 2023, acc: 2025-04-13.

[7] S. A. Lambert *et al.*, "The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation," *Nature genetics*, 2021.

[8] A. Manichaikul *et al.*, "Robust relationship inference in genome-wide association studies," *Bioinformatics*, 2010.

COLUMBIA
COLUMBIA UNIVERSITY
DEPARTMENT OF
BIOMEDICAL INFORMATICS

kirill.nikitin@columbia.edu
@ni_kirill

PRIVATE INFORMATION LEAKAGE FROM POLYGENIC RISK SCORES

Kirill Nikitin *Gamze Gürsoy*

NEW YORK GENOME CENTER
G²LAB

SUMMARY

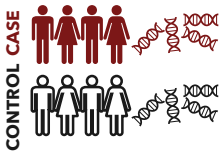
Polygenic Risk Scores (PRSs) are a popular clinical tool for the estimation of an individual's genetic susceptibility to diseases. We show that a profit-seeking actor, e.g., an insurance company, can recover the associated genotypes from publicly shared PRSs and use them to either de-anonymize individuals or infer sensitive traits of known patients.

BACKGROUND

G 0 1 0 1 2 0 0 1 0 0 1 0 0 0 0 2 2 0

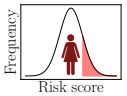
Genotypes as the difference count from the reference

GENOME-WIDE ASSOCIATION STUDIES



	SNP₁	SNP₂	SNP₃
Effect weight β	0.759	0.123	0.611

$$PRS = \frac{\sum_{i=1}^{N} \beta_i \times G_i}{Ploidy \times N}$$

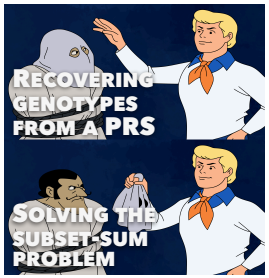


PRS example

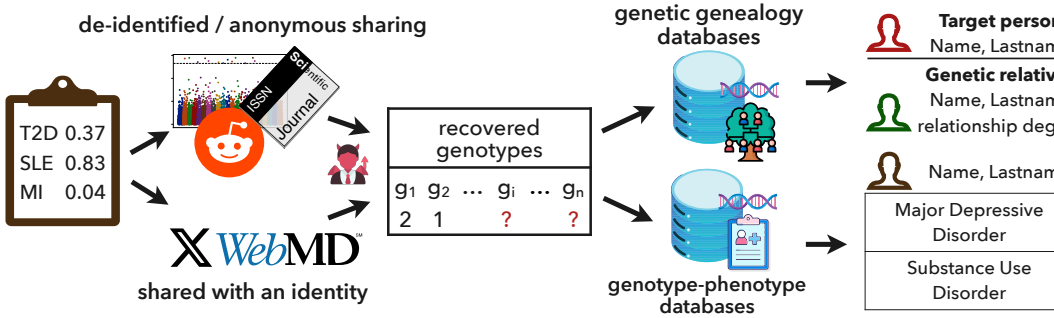
```
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000073
#trait_reported=Cervical cancer
#variants_number=10
#weight_type=log(OR)
rsID      chr_name chr_position effect  other.  effect_weight
rs3130196.    6      33063219     C       T     0.3576744442718159

rs3132461.    6      31480668     G       A     0.3074846997479607

rs2523557     6      31331257     G       A     -0.35065687161316933
```



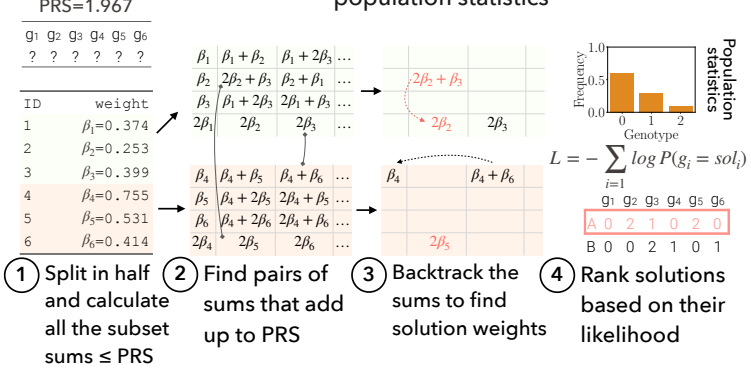
RISK SCENARIOS



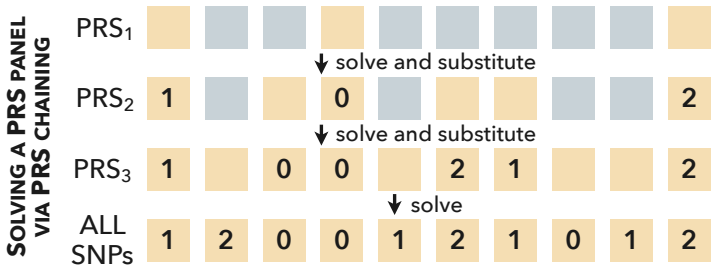
GENOTYPE RECOVERY

Given a PRS, we need to find how many times it includes each effect weight β

➡ Use dynamic programming and population statistics

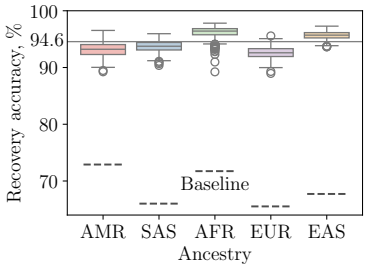


1. Split in half and calculate all the subset sums ≤ PRS
2. Find pairs of sums that add up to PRS
3. Backtrack the sums to find solution weights
4. Rank solutions based on their likelihood



EVALUATION

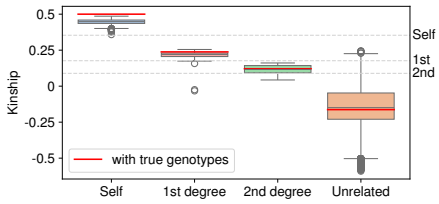
GENOTYPE RECOVERY ACCURACY



- Panel of 298 PRSs (2 to 49 SNPs) 4,821 SNPs total, 2,654 unique
- 2,535 genome samples from the 1000 Genomes Project
- We calculate PRSs and then attempt to recover genotypes
- 2,600 guessed SNPs on average

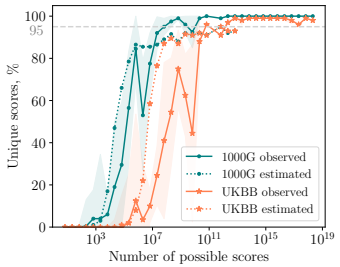
At least 454 out of 4,723 PRS models in the PGS Catalog are vulnerable to genotype recovery

GENEALOGY DATABASE SEARCH



- 68 samples with 1st or 2nd kins
- 100% accuracy identifying individuals and >80% accuracy identifying relatives

DATASET LINKING



In the UK Biobank dataset with 500K samples, a PRS model with 27 SNPs leads to 95% unique scores