PolyGenie: An Interactive Platform for Visualising Polygenic Risk Across Multidimensional Cohorts

Xavier Farré * (1,2), Mireia Gasco* (1,2), Natalia Blay (1,2,3) and Rafael de Cid (1,2)

1. Genomes for Life-GCAT Lab, CORE Program. Germans Trias i Pujol Research Institute (IGTP), Badalona, Spain.
2. Grup de Recerca en Impacte de Les Malalties Cròniques i Les Seves Trajectòries (GRIMTra), IGTP, Badalona, Spain.
3. Universitat de Barcelona (UB), Barcelona, Spain.

*Contributed equally

Corresponding Author

Rafael de Cid, PhD
Genomes for Life-GCAT lab. CORE Program
Germans Trias i Pujol Research Institute (IGTP)
08916 Badalona, Spain
Telf +34- 93 033 0542
e-mail: rdecid@igtp.cat

Biographical Note: The GCAT (Genomes for Life) team is based at the Institute for Health Science Research Germans Trias i Pujol (IGTP), Barcelona, Spain. The group leads a large prospective cohort study designed to explore the genetic, molecular, and environmental determinants of complex diseases.

Abstract

Summary:

Polygenic risk scores (PRS) quantify genetic liability to complex traits and diseases and have become increasingly powerful with the growth of large-scale GWAS summary statistics. Longitudinal population cohorts that integrate genetic, clinical, and lifestyle data provide a unique opportunity to explore PRS-trait associations within a unified framework. However, tools for systematic and interactive visualisation of these associations remain limited. We present PolyGenie, a web-based platform for phenome-wide exploration of PRS across diverse clinical and molecular phenotypes. Currently applied to 5,000 individuals from the GCAT cohort, PolyGenie integrates over 100 trait-specific PRS with a broad range of outcomes, including diseases, metabolite levels, lifestyle factors, and environmental exposures. The platform supports interactive PheWAS-style plots and outcome prevalence visualisation across PRS percentiles, enabling rapid interpretation of pleiotropy, genetic correlations, and population-level risk trends.

Availability and Implementation:

PolyGenie is freely available at https://polygenie.igtp.cat. The source code is available at (https://github.com/gcatbiobank/polygenie).

Contact: xfarrer@igtp.cat; rdecid@igtp.cat

Supplementary information:

Supplementary data are available at Bioinformatics online.

Introduction

Genome-wide association studies (GWAS) have transformed our understanding of the genetic basis of complex traits and diseases by identifying thousands of associated loci [1]. However, the highly polygenic nature of these traits, characterised by contributions from numerous genetic variants with small effect sizes, makes it difficult to use individual variants for accurate risk prediction in common disorders.

Polygenic risk scores (PRS) have emerged as powerful tools for quantifying genetic susceptibility to complex traits and diseases by aggregating the effects of numerous genetic variants [2]. PRS not only enhance the prediction of an individual's risk for specific traits but also enables researchers to explore genetic associations beyond a single trait. By examining how the polygenic risk score (PRS) for a given trait associates with a broad range of other phenotypes and diseases, researchers can uncover shared genetic architecture and pleiotropic effects, offering insight into the biological mechanisms that connect seemingly distinct conditions.

The growing availability of GWAS summary statistics, along with modern PRS techniques that go beyond selecting GWAS significant hits, and pipelines that ease their use [3], offer a powerful opportunity to deepen our understanding of the interplay between genetic susceptibility to complex traits and diseases and the wide array of variables captured in population cohorts. Particularly, cohorts linked to clinical records and enriched with multi-layer molecular data (e.g., genomics, metabolomics, exposomics) provide exceptional power to uncover biologically meaningful associations and translate genetic insights into actionable knowledge. To this end we present PolyGenie, an interactive web-based resource that enables exploration of associations between PRS for over 100 traits and a wide array of phenotypes, including diseases, metabolite levels, and lifestyle variables, based on data from the GCAT cohort [4], a large population-based cohort from Southern Europe. PolyGenie is set to expand its coverage to the entire GCAT cohort as additional genotype data and new incident cases become available, with dynamic medical updates planned annually, enhancing the platform's analytical capacity and overall impact. Additionally, PolyGenie supports exploration by providing intuitive visualisations that facilitate hypothesis generation. By deepening our understanding of the genetic basis of health and disease, the platform empowers future research and contributes to informed public health decision-making.

Several tools already support large-scale PRS and PheWAS analyses, each with strengths and limitations. The PRS Atlas [5] and SpPGS Atlas [6] serve as useful references—linking polygenic scores to traits in the UK population and showing population-specific distributions in Spain, respectively—but their reliance on traditional clumping and thresholding methods limits adaptability and excludes newer modelling approaches like Bayesian or machine learning techniques. While SpPGS allows disease-specific exploration, its heterogeneity in data origin, limited sample size, and lack of comprehensive phenotypic coverage reduce its utility for broader, hypothesis-driven research. Moreover, it does not integrate PheWAS visualisation alongside PRS results, limiting interpretability and exploratory potential. PheWAS R [7], while offering helpful tools for phenome-wide analyses, requires local setup and programming expertise, which may pose a barrier. In contrast, PolyGenie stands out as a practical, all-in-one platform that combines advanced PRS modelling with broad phenotypic data—ICD codes, Phecodes, metabolites, lifestyle, and environmental variables—within an accessible, browser-based platform. It supports integration within large cohorts like the UK Biobank and enables flexible PRS construction.

Overall, while existing tools contribute valuable insights, PolyGenie stands out as an integrated, user-friendly solution for exploring the genetic architecture of complex traits in population cohorts.

Methods

Individuals and variables.

The GCAT cohort is a comprehensive population-based study comprising 20,000 middle-aged individuals (40–65 years old) residing in Catalonia [4]. Recruitment began in 2014-2018, with periodic follow-ups ongoing. Participants benefit from over 15 years of linked clinical data via Electronic Health Records (EHRs) from Catalonia's public healthcare system, offering a rich resource of long-term health information. Genotype data are available for a subset of participants, enabling integrated analyses of genetic, clinical, and environmental factors influencing health and disease (http://igtp.cat/gcat).

In this study, we utilised available genetic data from a subset (n~5,000) of the GCAT cohort. Genome-wide genotyping was performed using the Infinium Expanded Multi-Ethnic Genotyping Array (MEGAEx) (Illumina, San Diego, California, USA). All included participants were of Iberian descent with White-Western European ancestry, as determined by self-reported data and principal component analysis (PCA) [8]. Imputation was carried out using the TOPMed reference panel [9], resulting in approximately 20 million unique autosomal variants. Only GCAT participants who passed stringent quality control measures were included (N=4,988). Variants with a minor allele frequency (MAF) greater than 0.001 and an imputation quality score (R^2) above 0.3 were retained for downstream analyses (N=17,203,152). Genotype data are accessible through the European Genome-phenome Archive (EGA) under accession ID EGAD00010001664.

General workflow

PRS calculation

A comprehensive collection of publicly available GWAS summary statistics, primarily derived from studies in individuals of European ancestry and selected for sample size and statistical power, was gathered from the GWAS Catalog [10] as well as other pertinent sources (see Supplementary Table 1 and the GWAS section on the web interface). Then, Polygenic risk scores (PRS) were computed for each set of GWAS summary statistics using the GenoPred pipeline [3] and the MegaPRS [11] method, and finally applied to all genetic dataset. PRS values were standardised as Z-scores for downstream analysis. Individuals were then stratified into deciles or quartiles for each calculated PRS.

PheWAS Analysis

We conducted a phenome-wide association study (PheWAS) for each polygenic risk score, defined as a continuous variable, assessing associations across multiple biological and contextual layers. These included diseases derived from electronic health records (EHRs), as well as molecular (metabolome), sociodemographic, lifestyle, and environmental variables collected directly within the cohort. Diseases were analysed using both ICD-10 codes and Phecodes; for ICD-10 data, codes were grouped at the three-character level. Phecodes were derived by mapping ICD-10 codes using standard procedures allowing for the aggregation of related diagnoses into clinically meaningful phenotypes [12].

Visualizing Prevalence Plots by Genetic Risk

To provide a more nuanced understanding of genetic liability, we calculated the relative prevalence of each disease at each polygenic risk score percentile. For each percentile, the

proportion of individuals with the disease was computed relative to the total number in that percentile. This fine-grained visualisation of how disease risk evolves with increasing genetic susceptibility supports the interpretation of pleiotropic effects and facilitates risk stratification.

Model implementation

Disease and phenotype outcomes were pre-analysed using linear or logistic regression models, adjusting for sex, age, and the first ten principal components to account for population stratification. Metabolite analyses included an additional covariate for chylomicrons to correct for non-fasting conditions during plasma collection (Carreras-Torres et al., 2024).

For each PRS, associations were tested separately across different outcome categories, including diseases (classified using ICD-10 codes and Phecodes), metabolite levels, and additional questionnaire, and urban exposome indicators (i.e., air pollution, built environment, and spatially-derived environmental exposures)[13]. To control for multiple testing, a Bonferroni-corrected significance threshold was applied independently within each outcome category. Specifically, the threshold was set to 0.05 divided by the number of outcomes tested per category for a given PRS. For example, when testing a PRS against all metabolites, significance was determined at 0.05 divided by the number of metabolites.

The underlying pipeline, depicted in Figure 1, is built to be scalable, facilitating the future integration of additional PRS and supporting expansion to the entire cohort as new data become available. All relevant data, including phenotypes, PRS values, covariates, and association results, are stored in a SQLite database to support efficient querying from the frontend.

User interface and visualisation

PolyGenie offers a user-friendly, interactive interface designed to facilitate the exploration of phenome-wide associations for selected polygenic risk scores (PRS). Users begin by selecting a PRS of interest from a searchable dropdown list. They can then choose to compare individuals in the top decile or quartile of the PRS distribution with those in the bottom decile or the rest of the cohort.

The primary visualisation is a PheWAS-style scatter plot, where outcomes are grouped by domain along the x-axis. The y-axis represents the signed $\log_{10}$(p-value), allowing users to assess both the strength and direction of associations. Hover functionality provides detailed summary statistics for each point, including effect sizes and p-values. Information about the GWAS source used to construct each PRS is also displayed below the main plot.

Upon selecting a specific phenotype, users are presented with a secondary plot showing the prevalence of that condition across PRS percentiles. This visualisation is accompanied by a data table summarizing the number of affected individuals in the cohort.

In addition to graphical outputs, results can be viewed in tabular format and exported for further analysis. Supplementary tabs give access to a searchable catalogue of all GWAS summary statistics used to generate PRS, including source references and trait descriptions. Another section provides cohort-level metrics such as age distribution and BMI, enabling users to contextualise genetic findings within the underlying population.

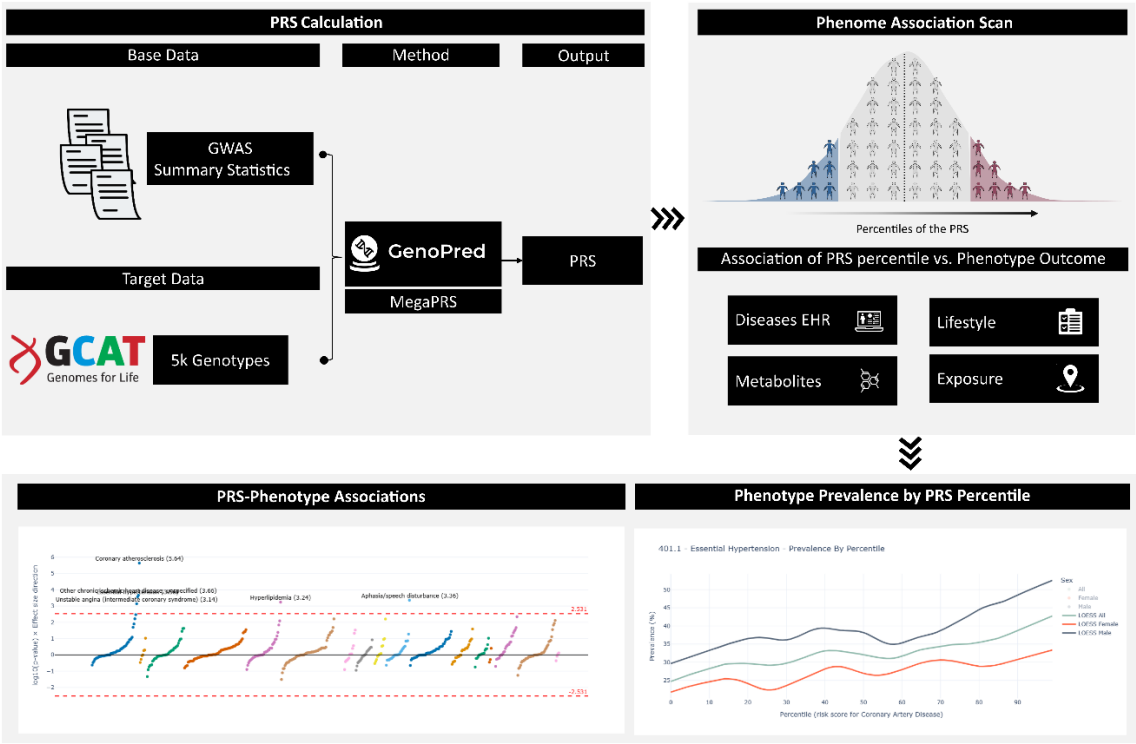

Figure 1. Overview of the PolyGenie platform: from PRS computation to phenome-wide visualization. *The upper panel illustrates the pipeline used to compute polygenic risk scores (PRS), integrating publicly available GWAS summary statistics and genotyped data from the GCAT cohort using the GenoPred pipeline and the MegaPRS method. The resulting standardized PRS values are used to conduct phenome-wide association scans (PheWAS). The lower panels show interactive visualization outputs available in the PolyGenie web application, including a PheWAS-like plot (left) that displays associations across diseases, metabolites, and other phenotypes, and a percentile-based prevalence plot (right) that shows how outcome prevalence varies with genetic liability, stratified by sex.*

Applying PolyGenie to Explore Sex Differences in Polygenic Risk for Frailty-Associated Traits

Fried frailty is strongly associated with both major depressive disorder (MDD) and overweight/obesity through shared behavioral, physiological, and social pathways [14]. Using genetic tools, specifically PolyGenie's interactive prevalence-by-percentile plots, we identified a clear dose–response relationship between higher Fried frailty PRS and increased prevalence of both major depressive disorder (MDD) and overweight/obesity (Figures 2a and 2b). The prevalence curve for depression was consistently higher in females across all PRS percentiles, mirroring established epidemiological trends—such as the approximately twofold higher 12-month prevalence of MDD in women compared to men [15]—and highlighting the tool's sensitivity to both biological and socially mediated sex differences. These patterns are in line with prior GWAS findings reporting genetic correlations between frailty, adiposity, and neuropsychiatric traits [16, 17]. Overall, our findings suggest that a greater genetic predisposition to frailty confers elevated susceptibility to both depressive illness and excess adiposity, consistent with a bidirectional interplay among these aging-related syndromes [18].

Although the current analysis is descriptive and PolyGenie is primarily designed for exploratory, hypothesis-generating use, it enables researchers to detect biologically relevant trends within harmonized, population-based cohorts like GCAT. The platform does not currently support direct covariate adjustment or the integration of heterogeneous external datasets, but its architecture facilitates scalable exploration across a wide range of phenotypes—including ICD codes, Phecodes, metabolites, and lifestyle variables—using consistent internal data from a single comprehensive cohort. When used alongside biobank-level infrastructures as GCAT or UKBiobanbank, PolyGenie offers a practical and intuitive environment for examining polygenic risk and trait associations across complex diseases and traits, that can support follow-up statistical analysis and model refinement through connected resources.

Figure 2. Prevalence of Clinical Outcomes Across Frailty PRS Percentiles. a) Prevalence of overweight and obesity (ICD-10: E66) increases approximately linearly with frailty polygenic risk score (PRS) percentile, with a comparable trend across sexes. b) Prevalence of major depressive disorder (ICD-10: F32) also rises with higher frailty PRS, with notably higher rates observed among females across all percentiles. These findings highlight the pleiotropic nature of genetic liability to frailty and its relationship with both metabolic and psychiatric conditions.

Conclusion

PolyGenie is a scalable, interactive platform that empowers basic researchers to explore shared genetic architecture across over 100 complex traits by linking polygenic risk scores (PRS) with a rich spectrum of clinical, molecular, lifestyle, and environmental phenotypes. Leveraging the deeply phenotyped GCAT cohort and a standardised PRS computation framework, PolyGenie enables phenome-wide association studies (PheWAS) alongside real-time visualisation of disease prevalence across PRS percentiles—a key innovation that facilitates intuitive

interpretation of genetic risk and pleiotropy. Unlike existing static PheWAS browsers, PolyGenie offers dynamic, longitudinal updates, making it a tool that evolves as new genomic and clinical data become available. Moreover, alongside publicly available PRS, user-defined PRS can be incorporated into the analysis pipeline to enable tailored exploration of the dataset.

By enabling systematic cross-trait comparisons, generating testable hypotheses, and uncovering potential biomarkers or pleiotropic effects, PolyGenie serves both research and translational goals. It supports precision medicine by making complex genetic information accessible through user-friendly visual analytics. Developed under open data principles, aligned with ELIXIR interoperability standards, and hosted within a robust, publicly maintained infrastructure, PolyGenie is designed for long-term sustainability as a FAIR-compliant resource supporting durable genomic research across European cohorts.

Key Points

- PolyGenie is an interactive web-based platform that links polygenic risk scores (PRS) for over 100 complex traits with a broad range of phenotypes—including diseases, metabolites, lifestyle, and environmental variables—to explore deeply phenotyped GCAT cohort, and that could be set up to other studies.
- The platform enables phenome-wide association studies (PheWAS) with real-time, percentile-based prevalence visualizations, supporting intuitive exploration of genetic risk and pleiotropy.
- Unlike existing PRS or PheWAS tools, PolyGenie integrates advanced PRS modeling, broad phenotypic coverage, and user-friendly interactive visualizations in a single browser-based interface.
- PolyGenie facilitates hypothesis generation, cross-trait comparisons, and identification of potential biomarkers, supporting both research and translational applications in precision medicine.
- The platform is scalable, FAIR-compliant, and designed for long-term sustainability, with dynamic updates as new genomic and clinical data become available.

Availability and Implementation

PolyGenie is freely accessible at https://polygenie.igtp.cat. To gain full access, and export data for downstream custom analysis, users are required to contact GCAT data owners.

The platform is implemented using the Plotly Dash framework for both backend and frontend, with an SQLite database for data storage and rapid querying. The application is deployed on a server maintained by the Institut Germans Trias i Pujol, ensuring high availability and responsiveness. While the web interface is designed to be compatible with all modern browsers and responsive across desktop and mobile devices, performance and rendering may vary slightly depending on specific browser implementations and mobile environments due to underlying limitations of the Dash framework. The service will remain publicly accessible for a minimum of two years following publication. Source code and documentation are available at (https://github.com/gcatbiobank/polygenie).

Competing interests

The authors declare no competing interests.

Funding

The Spanish Ministry of Science and Innovation (TED2021-130626B-I00).

Authors' contributions

All authors meet the criteria for authorship. RdC and XF conceived and designed the study. MG developed and implemented the platform, while NB led the implementation of the PheWAS analyses. RdC secured funding for the project. XF drafted the manuscript, with contributions from RdC and NB to the interpretation of results and overall framing of the work. All authors critically reviewed the manuscript for important intellectual content, approved the final version for publication, and agree to be accountable for all aspects of the work, ensuring that any questions related to its accuracy or integrity are appropriately investigated and resolved.

Acknowledgments

We express our sincere gratitude to all GCAT cohort study participants for their invaluable contributions. We also thank the team at the Blood and Tissue Bank, coordinated by Dr. Alonso Nogués and Dr. Grífols, for their essential support in sample collection. Special thanks are extended to the GCAT project investigators for their ongoing commitment and collaboration. Electronic Health Record (EHR) data were provided by the Catalan Agency for Quality and Health Assessment (PADRIS Program). A complete list of GCAT investigators can be found at www.genomesforlife.com.

References

1. Loos, R.J.F., *15 years of genome-wide association studies and no signs of slowing down.* Nat Commun, 2020. **11**(1): p. 5900.
2. Lewis, C.M. and E. Vassos, *Polygenic risk scores: from research tools to clinical instruments.* Genome Med, 2020. **12**(1): p. 44.
3. Pain, O., A. Al-Chalabi, and C.M. Lewis, *The GenoPred pipeline: a comprehensive and scalable pipeline for polygenic scoring.* Bioinformatics, 2024. **40**(10).
4. Obon-Santacana, M., et al., *GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia.* BMJ Open, 2018. **8**(3): p. e018324.
5. Richardson, T.G., et al., *An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome.* Elife, 2019. **8**.
6. Carmona, R., et al., *The Spanish Polygenic Score reference distribution: a resource for personalized medicine.* Eur J Hum Genet, 2025.
7. Carroll, R.J., L. Bastarache, and J.C. Denny, *R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment.* Bioinformatics, 2014. **30**(16): p. 2375-6.
8. Galvan-Femenia, I., et al., *Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort.* J Med Genet, 2018. **55**(11): p. 765-778.
9. Taliun, D., et al., *Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.* Nature, 2021. **590**(7845): p. 290-299.
10. Cerezo, M., et al., *The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity.* Nucleic Acids Res, 2025. **53**(D1): p. D998-D1005.
11. Zhang, Q., et al., *Improved genetic prediction of complex traits from individual-level data or summary statistics.* Nat Commun, 2021. **12**(1): p. 4192.
12. Wu, P., et al., *Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation.* JMIR Med Inform, 2019. **7**(4): p. e14325.

13. de Hoogh, K., et al., *A Europe-wide characterization of the external exposome: A spatio-temporal analysis.* Environ Int, 2025. **200**: p. 109542.
14. Feng, Z., et al., *Risk factors and protective factors associated with incident or increase of frailty among community-dwelling older adults: A systematic review of longitudinal studies.* PLoS One, 2017. **12**(6): p. e0178383.
15. Salk, R.H., J.S. Hyde, and L.Y. Abramson, *Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms.* Psychol Bull, 2017. **143**(8): p. 783-822.
16. Ye, Y., et al., *A genome-wide association study of frailty identifies significant genetic correlation with neuropsychiatric, cardiovascular, and inflammation pathways.* Geroscience, 2023. **45**(4): p. 2511-2523.
17. Liao, S.F., et al., *Associations of polygenic risks, depression, and obesity-related traits in Taiwan Biobank.* J Affect Disord, 2023. **320**: p. 397-403.
18. Zhu, J., et al., *Assessment of the bidirectional causal association between frailty and depression: A Mendelian randomization study.* J Cachexia Sarcopenia Muscle, 2023. **14**(5): p. 2327-2334.

Figure legends:

Figure 2. Overview of the PolyGenie platform: from PRS computation to phenome-wide visualization. The upper panel illustrates the pipeline used to compute polygenic risk scores (PRS), integrating publicly available GWAS summary statistics and genotyped data from the GCAT cohort using the GenoPred pipeline and the MegaPRS method. The resulting standardized PRS values are used to conduct phenome-wide association scans (PheWAS). The lower panels show interactive visualization outputs available in the PolyGenie web application, including a PheWAS-like plot (left) that displays associations across diseases, metabolites, and other phenotypes, and a percentile-based prevalence plot (right) that shows how outcome prevalence varies with genetic liability, stratified by sex.

Figure 2. Prevalence of Clinical Outcomes Across Frailty PRS Percentiles. a) Prevalence of overweight and obesity (ICD-10: E66) increases approximately linearly with frailty polygenic risk score (PRS) percentile, with a comparable trend across sexes. b) Prevalence of major depressive disorder (ICD-10: F32) also rises with higher frailty PRS, with notably higher rates observed among females across all percentiles. These findings highlight the pleiotropic nature of genetic liability to frailty and its relationship with both metabolic and psychiatric conditions.