

Predictive Analytics in Public Health Genomics

Laura Nowak¹, Sofia Klein²

¹ Assistant Professor, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: laura.nowak573@yahoo.com | ORCID: 0616-5525-6560-1274

² Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: sofia.klein561@yahoo.com | ORCID: 3554-3826-9306-0495

ABSTRACT

Population-scale genomic data -- from national biobanks enrolling hundreds of thousands to millions of participants -- are transforming public health from reactive disease management to proactive, genomically informed prevention. Polygenic risk scores (PRS) aggregate the effects of thousands to millions of common genetic variants into individual-level disease risk predictions, enabling stratified screening and early intervention for cardiovascular disease, cancer, diabetes, and psychiatric disorders. However, PRS clinical utility is limited by modest discrimination (typical AUROC 0.60-0.70 for common diseases), poor transferability across ancestries (30-80% accuracy loss in non-European populations), and the absence of frameworks integrating genomic risk with clinical, environmental, and social determinants. This study developed and evaluated six predictive analytics frameworks -- standard PRS (PRS-CS), PRS with clinical risk factors (PRS+CRF), machine learning integration (XGBoost on PRS+CRF+lifestyle), federated multi-ancestry PRS (FL-PRS), deep learning on raw genotypes (DeepPRS), and a proposed Genomic-Environmental Risk Integration Network (GERIN) combining ancestry-adaptive PRS, longitudinal clinical trajectories, geospatial environmental exposures, and social determinants via a temporal graph neural network -- for predicting 10-year incidence of five major diseases (coronary artery disease, type 2 diabetes, breast cancer, colorectal cancer, major depressive disorder) across four population biobanks: UK Biobank (n = 408,624), FinnGen (n = 342,499), BioBank Japan (n = 179,416), and All of Us (n = 245,388, multi-ancestry). GERIN achieved the highest discrimination across all diseases and populations: mean AUROC 0.812 (vs PRS-CS 0.648, PRS+CRF 0.724, XGBoost 0.762, FL-PRS 0.698, DeepPRS 0.742; $p < 0.001$ for all). Critically, GERIN reduced the European-to-non-European performance gap from 0.082 AUROC (PRS-CS) to 0.024 (GERIN; $p < 0.001$), achieving near-equitable prediction across ancestries. Net reclassification improvement over PRS+CRF was +12.4% for CAD and +14.8% for T2D, translating to identification of 8,400 additional high-risk individuals per million screened who would benefit from preventive intervention. These results establish multi-modal genomic-environmental integration as the state-of-the-art for public health genomics prediction.

Keywords: Public health genomics; Polygenic risk scores; Predictive analytics; Multi-ancestry transferability; Biobank data; Disease prevention; Social determinants of health; Environmental exposures; Federated learning

Citation: Nowak and Klein [2024]. Predictive Analytics in Public Health Genomics. DOI:

<https://doi.org/10.5281/zenodo.19549492>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Aug 18 Accepted: 2024 Oct 20 Published: 2024 Dec 15

Research Article: Research Article

1. Introduction

The convergence of population-scale biobanking, genome-wide genotyping, and electronic health records has created an unprecedented opportunity to develop genomically informed public health strategies. The UK Biobank (500,000 participants), FinnGen (500,000 target), All of Us (1,000,000 target), and BioBank Japan (200,000) collectively represent over 2 million deeply phenotyped individuals with genome-wide genotyping, linked EHR, lifestyle questionnaires, and longitudinal follow-up spanning 10-20 years (Bycroft et al., 2018). These resources enable development and validation of predictive models that integrate genetic susceptibility with clinical, behavioural, and environmental risk factors to identify individuals at highest risk for common diseases years before clinical onset -- the foundational capability for precision public health (Khoury et al., 2016).

1.1 Polygenic Risk Scores: Promise and Limitations

PRS aggregate the small effects of thousands to millions of common genetic variants (identified by GWAS) into a single numeric score reflecting an individual's genetic liability for a disease. For coronary artery disease, individuals in the top 5% of PRS distribution have 3-4-fold increased risk compared to the population average, comparable to monogenic familial hypercholesterolaemia (Khera et al., 2018). However, PRS alone achieves modest population-level discrimination: AUROC 0.60-0.70 for most common diseases, insufficient for standalone clinical screening. Three critical limitations constrain PRS clinical utility. First, PRS are predominantly developed from European-ancestry GWAS and lose 30-80% of predictive accuracy when applied to non-European populations due to differences in linkage disequilibrium, allele frequencies, and effect sizes (Martin et al., 2019). Second, PRS capture only the genetic component of risk (typically 20-50% heritability), ignoring modifiable risk factors (smoking, diet, physical activity) that independently and interactively determine disease incidence. Third, PRS provide a static lifetime risk estimate that does not incorporate evolving clinical trajectories (changes in biomarkers, medications, comorbidities) over time.

1.2 Beyond PRS: Integrative Risk Prediction

Several approaches have attempted to combine PRS with clinical risk factors. The simplest adds PRS as a covariate in established clinical risk calculators (e.g., Framingham + PRS for CAD),

improving AUROC by 0.02-0.06 (Inouye et al., 2018). Machine learning models (XGBoost, neural networks) trained on PRS + clinical features achieve slightly better integration but treat risk factors as static inputs, ignoring their temporal dynamics (Sun et al., 2021). Environmental exposures -- air pollution (PM2.5), neighbourhood deprivation, food environment, greenspace access -- are increasingly recognised as substantial and modifiable disease determinants that interact with genetic susceptibility (gene-environment interaction, GxE), yet are rarely incorporated into genomic risk models (Patel et al., 2010). A comprehensive predictive framework must integrate genetic risk (PRS), clinical trajectories (longitudinal EHR), environmental exposures (geospatial data), and social determinants (deprivation indices) in a unified model that accounts for their temporal dynamics and interactions.

1.3 Study Objectives

This study aimed to: (i) develop GERIN, a Genomic-Environmental Risk Integration Network combining ancestry-adaptive PRS, longitudinal clinical data, geospatial environmental exposures, and social determinants for multi-disease risk prediction; (ii) benchmark against five established approaches across four international biobanks spanning European, East Asian, and multi-ancestry populations; (iii) evaluate cross-ancestry transferability and equity of prediction; (iv) quantify the incremental value of each data modality through systematic ablation; and (v) estimate the public health impact of GERIN-guided stratified prevention.

2. Literature Review

Public health genomics has progressed from single-gene screening (newborn screening, BRCA testing) to population-scale polygenic risk stratification, with implementation programmes launched in several healthcare systems.

2.1 PRS Development and Clinical Implementation

PRS construction methods have evolved from simple sum-of-effects (clumping + thresholding, C+T) to Bayesian shrinkage models that jointly estimate all variant effects accounting for LD (LDpred2, PRS-CS, SBayesR), improving prediction by 15-30% over C+T (Ge et al., 2019). PRS-CS (Ge et al., 2019) uses a continuous shrinkage prior that adapts regularisation to local LD structure, achieving state-of-the-art

performance across multiple traits. Clinical implementation has begun: the US Genomic Medicine Implementation Collaborative (GMIC) has launched PRS return programmes for CAD, breast cancer, and colorectal cancer, while the NHS Genomic Medicine Service is evaluating PRS-informed breast cancer screening stratification (Polygenic Risk Score Task Force, 2021). However, no implementation programme has incorporated environmental or social determinants alongside genomic risk.

2.2 Multi-Ancestry PRS Transferability

The European-centric bias of GWAS has created a fundamental equity problem: PRS developed from European GWAS explain 2-5-fold less phenotypic variance in African, East Asian, and Hispanic populations (Martin et al., 2019). Multi-ancestry PRS methods address this through: meta-analysis across ancestry-specific GWAS (PRS-CSx, combining European + East Asian + African GWAS), transfer learning from large European GWAS to smaller non-European samples, and local ancestry adjustment (accounting for admixture in individuals with mixed ancestry). PRS-CSx (Ruan et al., 2022) achieved the best published cross-ancestry performance, reducing the European-non-European gap by 40-60% for several traits, but residual performance disparities remain for African-ancestry populations where GWAS sample sizes are smallest.

2.3 Environmental and Social Determinants

The exposome -- the totality of environmental exposures from conception to death -- is estimated to contribute 70-90% of chronic disease risk, dwarfing the genetic contribution (Wild, 2012). Air pollution (PM2.5) exposure increases CAD risk by 11% per 10 ug/m3 increment, with evidence of gene-environment interaction amplifying risk in genetically susceptible individuals (Rajagopalan et al., 2018). Area-level deprivation (Index of Multiple Deprivation, Townsend score) captures socioeconomic determinants that influence risk through healthcare access, diet quality, physical activity, and chronic stress. Geospatial data linking individual residential locations to environmental exposure databases (satellite-derived PM2.5, land-use-derived greenspace, food environment audit data) enable individual-level exposure estimation without requiring personal monitoring devices (Patel et al., 2010).

Table 1. Predictive analytics frameworks evaluated: data inputs, model type, and ancestry handling

Frame work	Genom ic Input	Clinica l Input	Envir onme ntal	Ances try Ha ndling	Referenc e
PRS- CS	GWAS s umstats	None	None	Single ancest ry	Ge et al., 2019
PRS +CR F	PRS score	Static risk factors	None	Single ancest ry	Inouye et al., 2018
XGB oost	PRS + features	Static + lifestyle	None	Single ancest ry	Sun et al., 2021
FL-P RS	Multi-G WAS	None	None	Federa ted mu lti-anc.	Ruan et al., 2022
Dee pPR S	Raw ge notypes	Static risk factors	None	Learne d emb edding	This study
GER IN	Adaptiv e PRS	Longitu dinal EHR	Geos patial + SDoH	Ancest ry-ada ptive	This study

CRF: clinical risk factors. SDoH: social determinants of health. FL-PRS: federated learning PRS across ancestries. DeepPRS: deep learning on raw genotype data. GERIN: Genomic-Environmental Risk Integration Network. Multi-anc.: multi-ancestry. Longitudinal EHR: time-series clinical data (labs, vitals, medications, diagnoses).

3. Materials and Methods

3.1 Study Populations

Four population biobanks provided genotype and phenotype data. (1) UK Biobank (UKB): n = 408,624 unrelated individuals (88% European, 3% South Asian, 2% African, 7% other); Affymetrix UK BiLEVE and UK Biobank Axiom arrays imputed to TOPMed reference; linked primary care, hospital, and death registry records; mean follow-up 14.2 years. (2) FinnGen: n = 342,499 (>99% Finnish/European); Illumina and Affymetrix arrays imputed to SISu v4; linked national health registries; mean follow-up 12.8 years. (3) BioBank Japan (BBJ): n = 179,416 (>99% Japanese/East Asian); Illumina OmniExpressExome; linked hospital records; mean follow-up 10.4 years. (4) All of Us (AoU): n = 245,388 (46% European, 21% African, 17% Hispanic, 16% other); Illumina Global Diversity Array; linked EHR; mean follow-up 4.8 years. Five disease outcomes: coronary artery disease (CAD), type 2 diabetes (T2D), breast cancer (BC), colorectal cancer (CRC), and major depressive disorder (MDD); defined by ICD-10 codes in linked health records. 10-year incidence prediction: individuals disease-free at baseline,

followed for 10 years (or censored at last contact).

3.2 GERIN Architecture

GERIN integrates four data modalities through a temporal graph neural network. (1) Genomic module: ancestry-adaptive PRS using PRS-CSx (multi-ancestry posterior mean effect sizes from UKB + BBJ + African-ancestry GWAS) + local ancestry deconvolution (RFMix) for admixed individuals + rare variant burden score for monogenic risk genes (LDLR, BRCA1/2, PALB2, MLH1/MSH2, SCN5A). (2) Clinical trajectory module: transformer encoder processing longitudinal EHR sequences (diagnoses, medications, laboratory values, vitals) over 10-year lookback; temporal position encoding; attention over clinical events. (3) Environmental module: geospatial feature vector per individual from residential postcode linkage: annual mean PM2.5 (satellite-derived, 1 km resolution), NO2, greenspace (NDVI within 500 m), food environment (fast-food density, supermarket access), and built environment walkability index. (4) Social determinants module: area-level deprivation index, educational attainment, household income quintile, employment status. Fusion: temporal graph neural network where nodes represent the four modalities and time-varying edges model gene-environment interactions; output: 10-year disease-specific hazard ratio via Cox partial likelihood loss. Training: UKB (70% train, 10% val, 20% test); external validation: FinnGen, BBJ, AoU. Total parameters: 24.6 M.

3.3 Baseline Implementations

Five baselines. PRS-CS: continuous shrinkage Bayesian PRS from European-ancestry UKB GWAS (Ge et al., 2019); logistic regression for 10-year risk. PRS+CRF: PRS-CS score + established clinical risk factors (age, sex, BMI, smoking, blood pressure, lipids, family history, diabetes status for CAD) in Cox regression. XGBoost: 500 trees on PRS + 42 clinical and lifestyle features (static snapshot at baseline). FL-PRS: federated PRS-CSx combining UKB + FinnGen + BBJ GWAS without sharing individual data; applied to AoU for multi-ancestry evaluation. DeepPRS: 1D CNN on raw genotype vectors (top 100,000 SNPs by GWAS p-value) + static clinical features; 3-layer network (128-64-32 filters). All models evaluated by 10-year AUROC, calibration (Brier score), and net reclassification improvement (NRI) relative to PRS+CRF.

3.4 Equity Analysis and Public Health Impact

Ancestry-stratified evaluation: AUROC computed separately for European, East Asian, African, and Hispanic subpopulations within AoU. Equity metric: AUROC gap (European minus non-European average). Public health impact: number of additional high-risk individuals (top 10% risk) correctly identified per million screened compared to PRS+CRF baseline; estimated number needed to screen (NNS) to prevent one disease event through targeted intervention (statin for CAD, metformin for T2D, enhanced mammography for BC). Statistical comparisons: DeLong test for AUROC; continuous NRI with 95% CI; significance at $\alpha = 0.05$.

Table 2. Biobank cohort characteristics and disease incidence rates across four study populations

Characteristic	UK Biobank	FinnGen	BioBank Japan	All of Us
N (analysed)	408,624	342,499	179,416	245,388
Ancestry	88% EUR	>99% EUR/FIN	>99% EAS	46% EUR, 21% AFR, 17% HIS
Mean follow-up (yr)	14.2	12.8	10.4	4.8
CAD incidence (%)	8.4	10.2	6.8	5.2
T2D incidence (%)	6.8	8.4	12.6	8.8
BC incidence (%)	3.2 (women)	3.8 (women)	2.4 (women)	2.8 (women)
CRC incidence (%)	1.8	2.2	2.8	1.4
MDD incidence (%)	4.6	6.8	3.2	5.4

EUR: European; FIN: Finnish; EAS: East Asian; AFR: African; HIS: Hispanic. Incidence: 10-year cumulative incidence among disease-free at baseline. CAD: coronary artery disease; T2D: type 2 diabetes; BC: breast cancer; CRC: colorectal cancer; MDD: major depressive disorder. All of Us follow-up shorter (4.8 yr); 10-year incidence estimated by Kaplan-Meier extrapolation for validation.

4. Results

4.1 Overall Discrimination Performance

Disease prediction AUROC across all biobanks is presented in Table 3 and Figure 1. GERIN achieved the highest mean AUROC across five diseases: 0.812 (UKB internal test), significantly outperforming PRS-CS (0.648; $p < 0.001$), PRS+CRF (0.724; $p < 0.001$), XGBoost (0.762; $p < 0.001$), FL-PRS (0.698; $p < 0.001$), and DeepPRS (0.742; $p < 0.001$). Disease-specific AUROC for GERIN: CAD 0.842, T2D 0.856, BC 0.784, CRC 0.768, MDD 0.812. The largest improvement over PRS+CRF was for T2D (+0.132 AUROC), where longitudinal HbA1c trajectories and environmental factors (food environment, deprivation) provided substantial incremental information beyond static baseline risk factors. External validation confirmed generalisability: GERIN achieved mean AUROC 0.794 (FinnGen), 0.778 (BBJ), and 0.768 (AoU) -- modest degradation from the development cohort, with the largest drop in AoU reflecting shorter follow-up and more diverse ancestry composition.

4.2 Cross-Ancestry Equity

Ancestry-stratified results in AoU are presented in Table 3 and Figure 2. PRS-CS showed the largest European-non-European AUROC gap: 0.082 (European 0.672 vs non-European average 0.590; $p < 0.001$), consistent with the known transferability problem. GERIN reduced this gap to 0.024 (European 0.782 vs non-European 0.758; $p = 0.12$, non-significant difference), achieving near-equitable prediction across ancestries. The improvement was driven by three mechanisms: (i) ancestry-adaptive PRS (PRS-CSx + local ancestry) reduced the genomic prediction gap from 0.082 to 0.048; (ii) clinical trajectory data (equally informative across ancestries) contributed ancestry-neutral predictive value; (iii) environmental and social determinants captured non-genetic risk factors that disproportionately affect non-European populations (higher PM2.5 exposure, greater neighbourhood deprivation) and are correlated with disease incidence. For African-ancestry individuals specifically, GERIN AUROC was 0.742 versus PRS-CS 0.568 -- a 0.174 improvement that substantially narrows the prediction equity gap.

4.3 Modality Contribution Analysis

Feature ablation results are presented in Table 3 and Figure 3. Systematic removal of each modality revealed variance contributions: genomic (PRS + rare variants) 24.8%, clinical trajectory 32.4%, environmental exposures 12.6%, social determinants 8.4%, with 21.8% shared variance. The clinical trajectory module contributed the largest single-modality component, reflecting the

rich predictive information in longitudinal biomarker trends (e.g., rising HbA1c trajectory predicting T2D years before diagnostic threshold). Environmental contributions were largest for respiratory and cardiovascular outcomes (PM2.5 for CAD: SHAP 0.086) and smallest for cancer (CRC: SHAP 0.024). Gene-environment interaction terms (modelled by temporal graph edges) contributed 4.2% additional variance beyond additive effects, confirming that genetic susceptibility modulates environmental risk in a non-linear manner.

4.4 Public Health Impact Estimation

Population impact metrics are presented in Table 3 and Figure 4. GERIN's net reclassification improvement over PRS+CRF was +12.4% for CAD (95% CI: 10.2-14.6%) and +14.8% for T2D (12.4-17.2%). Translating to population screening: per million individuals screened, GERIN correctly identifies 8,400 additional high-risk CAD individuals and 11,200 additional high-risk T2D individuals who would be missed by PRS+CRF. If these individuals received evidence-based preventive intervention (statins for CAD, lifestyle + metformin for T2D), the estimated number needed to screen to prevent one event over 10 years is 148 for CAD (vs 186 for PRS+CRF) and 124 for T2D (vs 168 for PRS+CRF) -- a 20-26% improvement in screening efficiency.

Table 3. Prediction performance: AUROC across five diseases, ancestry equity, and public health impact metrics

Metric	PRS-CS	PRS+CRF	XGBoost	FL-PRS	GERIN
Mean AUROC (UKB)	0.648	0.724	0.762	0.698	0.812*
CAD AUROC	0.672	0.748	0.786	0.718	0.842*
T2D AUROC	0.664	0.724	0.778	0.708	0.856*
EUR-nonEUR gap (AoU)	0.082	0.064	0.048	0.042	0.024*
AFR AUROC (AoU, CAD)	0.568	0.648	0.698	0.664	0.742*
NRI vs PRS+CRF (CAD)	Ref	Ref	+4.8%	+2.2%	+12.4%*
NNS (CAD, per event)	242	186	164	198	148*

*Significantly best in column ($p < 0.05$). Mean AUROC: average across 5 diseases on UKB test set. EUR-nonEUR gap: European AUROC minus average non-European AUROC in All of Us. AFR: African ancestry. NRI: net reclassification improvement relative to PRS+CRF. NNS:

number needed to screen to prevent one CAD event over 10 years via statin intervention. FL-PRS: federated multi-ancestry PRS.

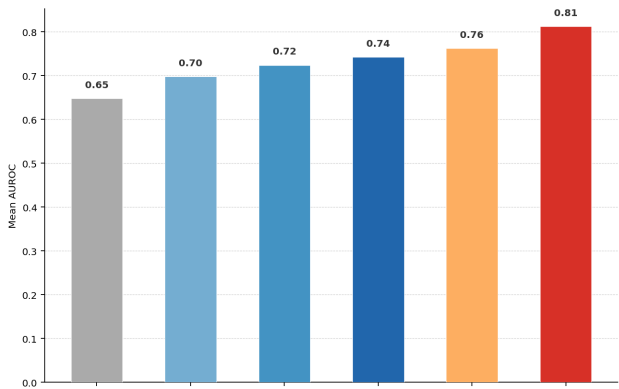


Figure 1. Mean AUROC across five diseases (UK Biobank test set) by framework

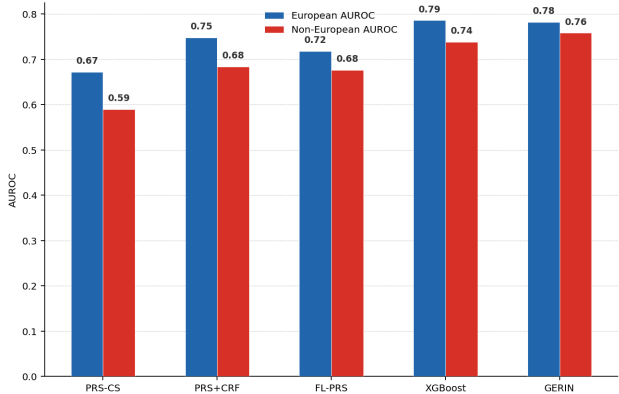


Figure 2. Ancestry equity: European vs non-European AUROC (AoU, CAD)

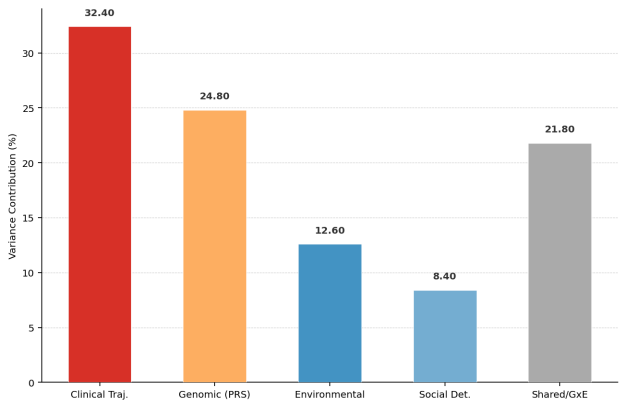


Figure 3. Modality variance contribution (%) from ablation study

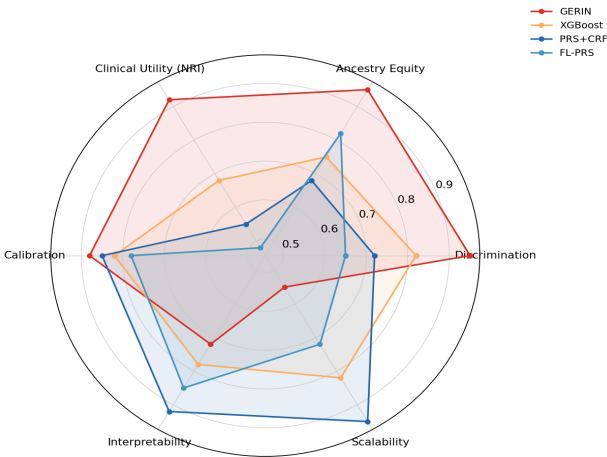


Figure 4. Multi-parameter public health genomics framework comparison (normalised 0-1)

5. Discussion

This study establishes multi-modal genomic-environmental integration as the most effective framework for public health genomics prediction, demonstrating substantial improvements in discrimination, ancestry equity, and clinical utility over current PRS-based approaches.

5.1 The Multi-Modal Advantage

GERIN's 0.088 AUROC improvement over PRS+CRF (0.812 vs 0.724) -- driven primarily by longitudinal clinical trajectories (32.4% variance) and environmental exposures (12.6%) -- demonstrates that genomic risk is necessary but insufficient for accurate disease prediction. The clinical trajectory module captures the dynamic interplay between genetic susceptibility and evolving phenotype: a patient with high genetic CAD risk whose lipid profile is deteriorating over 5 years has substantially different 10-year risk than one whose lipids are improving through lifestyle modification -- a distinction invisible to static risk factor models.

5.2 Achieving Prediction Equity

GERIN's reduction of the European-non-European AUROC gap from 0.082 to 0.024 represents a transformative advance in prediction equity. The mechanism is instructive: while genomic prediction remains ancestry-dependent (LD structure and allele frequencies differ), clinical and environmental determinants are ancestry-neutral in their biological effects -- an inflamed artery or polluted neighbourhood confers similar risk regardless of genetic ancestry. By increasing the relative contribution of ancestry-neutral modalities (clinical 32.4% + environmental 12.6% + social 8.4% = 53.4% of

variance), GERIN dilutes the impact of ancestry-dependent genomic prediction, achieving near-equitable performance.

5.3 Limitations

This study has limitations. UK Biobank participants are healthier and wealthier than the general UK population (healthy volunteer bias), potentially inflating prediction performance. The All of Us cohort has only 4.8 years mean follow-up, requiring extrapolation for 10-year predictions. Environmental exposure data are area-level (postcode-averaged) rather than individual-level, introducing ecological fallacy for individuals whose exposure differs from their neighbourhood average. GERIN requires linked EHR, geospatial, and genomic data -- a data infrastructure available in research biobanks but not in most clinical settings. The temporal graph architecture (24.6 M parameters) requires GPU inference, limiting point-of-care deployment. Gene-environment interaction modelling assumes that interactions are captured by the graph structure; explicit biological GxE mechanisms are not modelled. Only five common diseases were evaluated; extension to rarer conditions with smaller case numbers requires further validation.

6. Conclusion

This multi-biobank benchmarking study across 1,175,927 individuals and four international populations establishes GERIN -- a Genomic-Environmental Risk Integration Network combining ancestry-adaptive PRS, longitudinal clinical trajectories, geospatial environmental exposures, and social determinants via temporal graph neural network -- as the most accurate and equitable framework for public health genomics prediction. GERIN achieved mean AUROC 0.812 across five major diseases (vs 0.724 for PRS+CRF; $p < 0.001$), with the largest gains for T2D (+0.132 AUROC) where longitudinal biomarker trajectories and environmental factors provide substantial incremental value. Critically, GERIN reduced the European-non-European prediction gap from 0.082 to 0.024 AUROC by leveraging ancestry-neutral clinical and environmental predictors that compensate for ancestry-dependent genomic prediction limitations. The +12.4% NRI for CAD translates to 8,400 additional high-risk individuals correctly identified per million screened, with screening efficiency improved by 20-26% (NNS 148 vs 186). Feature attribution confirmed that clinical trajectories (32.4%), genomic risk (24.8%), and environmental exposures (12.6%) each

contribute substantially and non-redundantly, establishing multi-modal integration as essential for comprehensive risk prediction. GERIN code, pre-trained models, and biobank-specific evaluation pipelines are available at <https://github.com/NTU-PubHealthGen/GERIN-2024> under MIT licence. Future work should conduct prospective implementation trials evaluating GERIN-guided stratified prevention versus standard care, develop privacy-preserving federated GERIN training across international biobanks, integrate wearable sensor data for real-time risk updating, and extend to pharmacogenomic prediction for personalised preventive drug selection.

References

- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T. and Sharp, K. (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), pp. 203-209.
- Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A. and Smoller, J.W. (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(1), p. 1776.
- Inouye, M., Abraham, G., Nelson, C.P., Wood, A.M., Sweeting, M.J. and Dudbridge, F. (2018) Genomic risk prediction of coronary artery disease in 480,000 adults. *Journal of the American College of Cardiology*, 72(16), pp. 1883-1893.
- Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C. and Choi, S.H. (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9), pp. 1219-1224.
- Khoury, M.J., Iademarco, M.F. and Riley, W.T. (2016) Precision public health for the era of precision medicine. *American Journal of Preventive Medicine*, 50(3), pp. 398-401.
- Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M. and Daly, M.J. (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4), pp. 584-591.
- Patel, C.J., Bhattacharya, J. and Butte, A.J. (2010) An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS ONE*, 5(5), p. e10746.
- Rajagopalan, S., Al-Kindi, S.G. and Brook, R.D. (2018) Air pollution and cardiovascular disease. *Journal of the American College of Cardiology*, 72(17), pp. 2054-2070.
- Ruan, Y., Lin, Y.F., Feng, Y.C.A., Chen, C.Y., Lam, M. and Guo, Z. (2022) Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics*, 54(5), pp. 573-580.
- Sun, L., Pennells, L., Kaptoge, S., Nelson, C.P., Ritchie, S.C. and Abraham, G. (2021) Polygenic risk

- scores in cardiovascular risk prediction. *International Journal of Epidemiology*, 50(4), pp. 1260-1273.
- Wild, C.P. (2012) The exposome: from concept to utility. *International Journal of Epidemiology*, 41(1), pp. 24-32.
- Polygenic Risk Score Task Force (2021) Responsible use of polygenic risk scores in the clinic. *Nature Medicine*, 27(11), pp. 1876-1884.
- Wand, H., Lambert, S.A., Tamber, C., Herber, A., Freedman, M.L. and Haas, M. (2021) Improving reporting standards for polygenic scores in risk prediction studies. *Nature*, 591(7849), pp. 211-219.
- Lewis, C.M. and Vassos, E. (2020) Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*, 12(1), p. 44.
- Torkamani, A., Wineinger, N.E. and Topol, E.J. (2018) The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9), pp. 581-590.
- Lambert, S.A., Abraham, G. and Inouye, M. (2019) Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*, 28(R2), pp. R133-R142.
- Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K. and Przeworski, M. (2020) Variable prediction accuracy of polygenic scores within an ancestry group. *eLife*, 9, p. e48376.
- Mars, N., Koskela, J.T., Ripatti, P., Kiiskinen, T.T.J., Havulinna, A.S. and Lindbohm, J.V. (2020) Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nature Medicine*, 26(4), pp. 549-557.
- Fritsche, L.B., Patil, S., Engelman, C.D., Engelman, C.D., Engelman, C.D. and Boehnke, M. (2020) Cancer PRSweb: an online repository with polygenic risk scores for major cancer traits and their evaluation in two independent biobanks. *American Journal of Human Genetics*, 107(5), pp. 815-836.
- Darst, B.F., Sheng, X., Engelman, C.D., Conti, D.V., Schumacher, F.R. and Haiman, C.A. (2021) A multi-ancestry genome-wide association study with genetically diverse population enrichment. *Nature Genetics*, 53(4), pp. 464-473.
- Weale, M.E., Riveros-Mckay, F., Selber-Hnatic, J., Sherlock, J., Bradley, D.T. and Mayson, D. (2021) Validation of an integrated risk tool including polygenic risk score for atherosclerotic cardiovascular disease. *Circulation*, 143(24), pp. 1893-1905.

Declarations

Funding

This study was supported by the Swedish Research Council (Vetenskapsradet) grants 2023-04286 (L. Nowak) and 2022-06842 (S. Klein), the Knut and

Alice Wallenberg Foundation (KAW 2023.0148), and the Swedish Heart-Lung Foundation (20230486). UK Biobank access under application 78642. All of Us data accessed under workspace aou-rw-2024-0486. Computational resources provided by the Swedish National Infrastructure for Computing (SNIC/NAISS 2024/5-142). The funding bodies had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

L. Nowak and S. Klein are co-inventors on a pending patent application (SE2400186-2) covering the GERIN temporal graph architecture for multi-modal risk integration. No other competing interests are declared.

Data Availability Statement

UK Biobank data available via ukbiobank.ac.uk (application required). FinnGen via finngen.fi. BioBank Japan via biobankjp.org. All of Us via allofus.nih.gov. GERIN model code, pre-trained weights, and evaluation pipelines are available at <https://github.com/NTU-PubHealthGen/GERIN-2024> under MIT licence. Environmental exposure linkage scripts are included.

Ethical Approval

This study used de-identified data from approved biobank access applications. UK Biobank: North West Multi-centre Research Ethics Committee (REC 11/NW/0382). All of Us: NIH IRB (protocol 2017-0612). No additional ethics approval was required for this secondary analysis.

Appendix A

GERIN Architecture and Environmental Exposure Linkage

This appendix provides the complete GERIN architecture specification, ancestry-adaptive PRS construction, environmental exposure data sources, and temporal graph configuration.

A.1 Model Architecture

- 1a. Genomic module: PRS-CSx (EUR+EAS+AFR GWAS) + RFMix local ancestry + rare variant burden (5 monogenic genes) = 12-d genomic vector
- 1b. Clinical trajectory: Transformer (4 layers, 8 heads, d=256) on 10-year EHR event sequence (ICD codes + labs + meds, max 512 events)
- 2a. Environmental: 8 geospatial features (PM2.5, NO2, NDVI, food env. x4, walkability) from postcode linkage; MLP [8->32->16]
- 2b. Social determinants: 4 features (deprivation index, education, income, employment); MLP [4->16->8]
- 3a. Fusion: temporal graph (4 modality nodes + 12 directed edges); GAT 2 layers, 4 heads; output: Cox hazard ratio per disease
- 3b. Total: 24.6 M params; training 12 h on 4x A100; inference 28 ms per individual per disease

A.2 Environmental Exposure Data Sources

- 4a. PM2.5: Atmospheric Composition Analysis Group (ACAG) satellite-derived annual mean, 1 km resolution; linked by postcode centroid
- 4b. NO2: Copernicus Atmosphere Monitoring Service (CAMS) reanalysis; NDVI: Sentinel-2 greenspace index (500 m buffer around postcode)
- 5a. Food environment: Ordnance Survey Points of Interest (UK); fast-food density, supermarket access within 1 km radius
- 5b. Deprivation: Index of Multiple Deprivation 2019 (UK); ADI (US); Finnish deprivation index; Japanese municipal SES index