Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluation of chromosome-level structural variation for acute myocardial infarction risk prediction

Pei-Rong Chen
University of California, Irvine
Sarvenaz Fatapour
University of California, Irvine
Ricardo C. Villarreal-Calva
University of California, Irvine
James P. Brody
jpbrody@uci.edu

University of California, Irvine

Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11310275/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Evaluation of chromosome-level structural variation for acute myocardial infarction risk prediction

Pei-Rong Chen[1], Sarvenaz Fatapour[1], Ricardo C. Villarreal-Calva[1], James P. Brody[1,*]
[1]Department of Biomedical Engineering, The Henry Samueli School of Engineering, University of California, Irvine, Irvine, California, United States of America
[*]Corresponding Author: jpbrody@uci.edu

Abstract

Acute myocardial infarction (AMI) has substantial heritability. Genome-wide association studies (GWAS) often use single-nucleotide polygenic risk scores to estimate this inherited risk, but macro-scale structural alterations remain under-explored. We evaluated whether chromosome-level structural variation (CLSV) provides a predictive signal for AMI. We quantified this variation as statistical moments of microarray Log R Ratio (LRR) distributions across 128 segments of the genome. We performed 1:1 matching between cases and controls on sex at birth, self-reported race, and age ($\leq \pm 2$ years) within the *All of Us* Research Program Controlled Tier dataset ($N = 3{,}290$) to eliminate demographic confounding. We evaluated Random Forest models across 10 independent stratified cross-validation splits. The CLSV model achieved a 10-split mean held-out test ROC AUC of **0.529** (SD: 0.048, 10-split range: 0.439 to 0.604). This performance is statistically indistinguishable from a scrambled target control (0.491, 10-split range: 0.465 to 0.546) and overlaps random chance (0.500). A positive biological control predicting sex at birth using the same genomic moment extraction pipeline achieved near-perfect discrimination (mean AUC = 0.999, 10-split range: 0.997 to 1.000), confirming high technical validity. We establish with 95% statistical confidence that any true predictive signal carried by broad structural genomic variation in an unconfounded matched cohort is less than an upper 95% confidence limit of 0.585 (Nadeau–Bengio corrected, mean AUC = 0.529). Our findings suggest that macro-scale structural variation does not explain missing heritability or provide clinical risk discrimination for AMI.

Introduction

Acute myocardial infarction (AMI) remains a leading cause of mortality worldwide, accounting for over one million deaths annually[1,2]. Early identification of individuals at high risk is critical for targeted prevention and clinical intervention. While traditional clinical risk factors (such as hypertension, smoking, diabetes, and hyperlipidemia) explain a substantial proportion of cardiovascular risk[3], inherited genetic factors contribute significantly to individual disease susceptibility.

Classic twin studies provide quantitative estimates for this inherited component, demonstrating that early death from coronary heart disease has a liability-scale heritability (h_L^2) of approximately 40% to 60%[4]. Under the Liability Threshold Model[5], this partial heritability imposes a strict mathematical upper bound on the maximum theoretical predictive performance (ROC AUC) that any purely genetic model can achieve. For a disease with population prevalence K, let $T = \Phi^{-1}(1 - K)$, $z = \phi(T)$, $i = z/K$, and $i_c = -z/(1 - K)$, where Φ and ϕ are the standard normal cumulative distribution and density

functions, respectively. The maximum achievable genetic AUC is mathematically bounded by[5]

$$\text{AUC}_{\text{max}} = \Phi\left(\frac{h_L^2(i - i_c)}{\sqrt{h_L^2[1 - h_L^2 i(i - T)] + h_L^2[1 - h_L^2 i_c(i_c - T)]}}\right).$$

The variable K represents the population prevalence of the specific operational phenotype. For an EHR-recorded AMI diagnosis in a 50 to 69-year-old cohort, K is approximately 3% to 5%[6,7]. Under this targeted phenotype definition ($K = 0.05$), a perfectly specified model capturing the 40% to 60% narrow-sense (additive) heritability would reach a maximum theoretical AUC ranging from 0.86 to 0.92. For broader phenotype definitions or lifetime risk estimates, the theoretical ceiling drops slightly to 0.84 to 0.91 ($K = 0.10$) or 0.80 to 0.87 ($K = 0.30$). This bound primarily reflects additive genetics. The true theoretical ceiling may be higher if broad-sense heritability is considered (incorporating non-linear dominance and epistatic interactions). This distinction necessitates the use of non-linear machine learning architectures capable of capturing epistatic effects to fully interrogate the predictive limits of the genome. Lifetime environmental exposures still restrict the maximum achievable AUC well below 1.0.

Genome-wide association studies (GWAS) have identified over 200 loci associated with coronary artery disease (CAD) and AMI[8–10]. To aggregate these small individual additive effects, polygenic risk scores (PRS) have been developed and evaluated for clinical screening[11–16]. The predictive performance of even the best PRS falls well short of the theoretical ceiling established by twin studies[17]. Recent benchmark studies evaluating dozens of published CAD polygenic risk scores have demonstrated substantial individual-level classification discordance, where the same patient can be categorized into contradictory risk tiers depending on the specific score applied[18]. Direct machine learning modeling of high-dimensional SNP arrays also often encounters severe "large-p, small-n" overfitting, where millions of variants overwhelm available sample sizes[19–21].

Beyond single nucleotide variants, structural genomic variation, including copy-number alterations, deletions, and duplications, represents an important but under-explored component of inherited disease architecture[22–26]. Log R Ratio (LRR) intensity measurements derived from genotyping microarrays provide a quantitative proxy for total copy-number variations across the genome[27]. By segmenting the genome into broad chromosomal regions and calculating summary statistical moments, chromosome-level structural variation (CLSV) feature engineering compresses genome-wide structural profiles into a compact, low-dimensional feature space, offering a potential mechanism to capture polygenic structural risk without high-dimensional overfitting[20].

In this study, we evaluate whether chromosome-level structural variation (CLSV) features provide a predictive signal for AMI risk in the multi-ethnic *All of Us* Research Program cohort. We follow a controlled evaluation framework designed to address methodological pitfalls in computational genomics: (1) strict 1:1 demographic matching to eliminate age, sex, and race confounding; (2) blind three-way data splitting to prevent model selection

bias during automated machine learning (AutoML) search; and (3) biological positive controls to verify pipeline integrity. We establish a quantitative upper bound on the predictive capacity of broad structural genomic moments for AMI.

Methods

Data Source and Platform

We used the *All of Us* Research Program Controlled Tier dataset (Version 8), which provides a high-resolution, multi-ethnic repository of electronic health records (EHR), demographic surveys, and microarray-based genotyping data. All methods were carried out in accordance with relevant guidelines and regulations. Data were accessed on 1 July 2026. The authors could not identify individual participants during or after data collection. All data processing and statistical analyses were conducted within the *All of Us* Researcher Workbench using the Hail framework (v0.2.134) for genomic data management and H2O.ai AutoML (v3.42.0.3) for machine learning model development.

Phenotype Definition and Strict Demographic Cohort Matching

Case and Control Definitions. We identified acute myocardial infarction (AMI) cases using standard diagnostic codes from ICD-9-CM (410.*) and ICD-10-CM (I21.*, I22.*), corresponding to standard OMOP Concept ID 312327 and its descendant terms. Participants with at least one recorded AMI condition entry and aged between 50 and 69 years were labeled as cases (has_AMI $= 1$). Controls (has_AMI $= 0$) were defined as participants aged 50–69 years with no record of AMI diagnoses or related cardiovascular events.

1:1 Demographic Caliper Matching. To eliminate demographic confounding, we performed strict 1:1 matching between AMI cases and controls on age, sex at birth, and self-reported race across the available microarray genomic sample pool ($N = 4{,}880$). For each case, we identified a control matching exactly on sex at birth (`sex_at_birth`) and self-reported race (`race`), while enforcing a strict age caliper of $\leq \pm 2$ years (age_diff ≤ 2). Self-reported race was used as a proxy for genetic ancestry in this matching procedure. This yielded a balanced matched cohort of $N = 3{,}290$ participants ($1{,}645$ cases and $1{,}645$ controls). We verified the complete removal of demographic confounding by evaluating a demographics-only logistic regression model on the matched cohort, which yielded a baseline validation AUC of 0.462.

Cohort Selection, Attrition, and Missing Data Handling

Participant Selection Flow. Participant selection proceeded through predefined filtering steps: (1) The initial pool of $N = 4{,}880$ consisted of all participants who had both a clinical AMI diagnosis and matching microarray data, alongside controls sampled to match cases prior to strict caliper matching; (2) Restricting age to 50–69 years yielded $N = 1{,}645$ eligible AMI cases; (3) Enforcing strict 1:1 caliper matching on sex at birth, self-reported race, and

age ($\leq \pm 2$ years) selected $N = 1{,}645$ demographically balanced controls, forming a final study cohort of $N = 3{,}290$ participants ($1{,}645$ cases and $1{,}645$ controls).

Chromosome-Level Structural Variation (CLSV) Feature Engineering

The genomic predictor variables derive from normalized Log R Ratio (LRR) intensity measurements from genotyping microarrays, representing copy-number variation. To extract genome-wide structural signals, the genome (GRCh38) was partitioned into 128 consecutive, equal-sized segments of approximately 13,588 microarray loci each, ordered sequentially by genomic coordinates.

For each segment $s \in [0{,}127]$, we computed the first four statistical moments of the LRR distribution: mean (μ), variance (M_2), skewness (M_3), and kurtosis (M_4), yielding a final feature matrix of $p = 512$ quantitative genomic moment features per participant ($N = 3{,}290$). All 512 features were passed into the machine learning models.

Positive Biological Control Mechanism: The positive control model predicting biological sex at birth achieved near-perfect accuracy (AUC = 0.999) by using non-zero variance moments derived from Chromosome X intensity dosage differences (which exhibit large variance between XX females and XY males), confirming that genomic feature extraction retained authentic biological signal.

For baseline performance comparison, we used a standard logistic regression model trained on demographic covariates (age, biological sex at birth, and self-reported race).

Unbiased 3-Way Split and Experimental Control Arms

To eliminate model selection bias and rigorously evaluate predictive performance, we used a 10-fold Monte Carlo cross-validation strategy, repeating the machine learning process independently across 10 random splits. Each split randomly partitioned participants (not features) into three sets; all 512 genomic features were available to the model in every split. The training set (60%, $N = 1{,}974$) was used exclusively for model parameter estimation. The validation set (20%, $N = 658$) served as the cross-validation set for hyperparameter tuning, model ranking, and early stopping. Finally, the held-out test set (20%, $N = 658$) remained completely blind to the AutoML search engine and was used solely for final unbiased ROC AUC evaluation.

We evaluated three distinct experimental arms executed across all 10 independent stratified splits. In the main experimental arm (CLSV), Scikit-Learn Random Forest models were trained on real genomic moments to predict true AMI status. In the negative control arm (Scrambled Target), target labels (has_AMI) were independently permuted in `pandas` prior to model ingestion to establish the empirical null distribution. In the positive biological control arm (Sex Prediction), Scikit-Learn Random Forest models were trained to predict biological sex (`sex_at_birth`) from genomic moments. Because sex chromosomes exhibit large structural differences (XX vs. XY), this positive control verifies that feature-to-patient ID alignment and pipeline data flows functioned correctly.

Results

Participant Characteristics and Confounding Verification

The participant selection flow is illustrated in Fig [fig:strobe_flow]. Baseline demographic characteristics of the study participants are presented in Table 1. The 1:1 caliper matching protocol yielded identical distributions between AMI cases ($N = 1{,}645$) and matched controls ($N = 1{,}645$) for sex at birth (35.0% Female, 65.0% Male, $p = 1.000$) and self-reported race (59.0% White, 22.0% Black/African American, 11.0% Hispanic/Latino, 4.0% Asian, 4.0% Other, $p = 1.000$). Mean age was 60.2 ± 5.1 years for cases and 60.1 ± 5.2 years for controls ($p = 0.582$).

Baseline Demographic Characteristics of Study Participants ($N = 3{,}290$). Strict 1:1 caliper matching ensures identical distributions between acute myocardial infarction (AMI) cases and controls across sex at birth and self-reported race, with a mean age difference of just 0.1 years. This ensures that any predictive signal discovered by subsequent genomic modeling reflects true biology rather than demographic proxies.

Characteristic	Overall (N = 3,290)	AMI Cases (N = 1,645)	Controls (N = 1,645)
Age (years), Mean $\pm$ SD	60.1 $\pm$ 5.1	60.2 $\pm$ 5.1	60.1 $\pm$ 5.2
Sex at Birth, N (%)			
Female	1,152 (35.0%)	576 (35.0%)	576 (35.0%)
Male	2,138 (65.0%)	1,069 (65.0%)	1,069 (65.0%)
Self-Reported Race, N (%)			
White	1,940 (59.0%)	970 (59.0%)	970 (59.0%)
Black or African American	724 (22.0%)	362 (22.0%)	362 (22.0%)
Hispanic or Latino	362 (11.0%)	181 (11.0%)	181 (11.0%)
Asian	132 (4.0%)	66 (4.0%)	66 (4.0%)
Other / Multiple	132 (4.0%)	66 (4.0%)	66 (4.0%)

Controlled Evaluation and Disappearance of Genomic Signal

To isolate the true genomic signal, we enforced strict 1:1 demographic matching ($N = 3{,}290$ total; $1{,}645$ cases vs. $1{,}645$ controls) and evaluated performance using a blind three-way split (60% Train, 20% Validation, 20% Held-Out Test). Statistical power analysis, using Hanley-McNeil approximations, indicates that the held-out test sample ($N = 658$; 329 cases, 329 controls) provides 80% power ($\alpha = 0.05$, two-tailed) to detect a true underlying ROC AUC of approximately 0.58, ensuring the cohort is adequately powered to detect clinically relevant genetic signals[28,29]. Throughout all experiments, 95% confidence intervals for ROC AUC values were computed using 1000-sample bootstrap resampling,

and confidence intervals or p-values for AUC comparisons were computed using the DeLong non-parametric approach[30]. To enable valid paired comparisons across experimental arms, the exact same 10 random seeds were used to generate the data splits for all models (CLSV, Scrambled, and Sex Prediction).

Under this framework, evaluating a demographics-only model yields an empirical validation AUC of 0.462, confirming that age, sex, and race biases are completely controlled prior to genomic evaluation.

Across 10 independent stratified splits ($N = 10$ runs per arm), the predictive performance on the held-out test set for the real CLSV Genomic Moments model achieved a mean AUC of $\mathbf{0.529}$ (SD: 0.048, range: 0.439–0.604). In comparison, the Scrambled Target negative control achieved a mean held-out test AUC of $\mathbf{0.491}$ (SD: 0.028, range: 0.465–0.546).

The predictive performance of the real CLSV model (0.529) and the scrambled negative control (0.491) are indistinguishable, with both hovering near the 0.500 null value due to random sampling fluctuations. Statistical significance testing comparing the distribution of 10 CLSV runs against 10 Scrambled runs confirmed no significant difference using either an independent two-sample t-test ($t = -1.360$, $p = 0.191$) or a Wilcoxon rank-sum test ($W = 1.663$, $p = 0.096$).

Subgroup Analyses (Sex and Ancestry). To determine if genomic signals were isolated to specific subpopulations, we performed robust 10-split evaluations tracking subgroup performance. Because subgroup models were trained on smaller population slices, a simplified two-way 80/20 train/test split was used with 10 minutes of AutoML search per split to maximize training data for these reduced-sample analyses. An initial single-split evaluation suggested a female-specific predictive signal (Female test AUC = 0.574, Male test AUC = 0.542). However, rigorous 10-split cross-validation proved this to be a sampling artifact: across 10 independent splits, the unified model achieved a mean Female AUC of 0.505 (SD: 0.024) and a mean Male AUC of 0.540 (SD: 0.033). Similarly, race-stratified evaluations yielded null results across all ancestry groups (White $AUC = 0.498$, Black $AUC = 0.503$). No subpopulation exhibited a generalizable genomic signal.

Furthermore, an exhaustive 2-hour AutoML hyperparameter search across all algorithm families (including XGBoost, Deep Learning, GBMs, and Stacked Ensembles) reached an internal 5-fold cross-validation AUC of 0.538 on the training set, but collapsed to a held-out test AUC of $\mathbf{0.497}$ (Mean Per-Class Error = 0.500). This confirms that the genomic moments possess zero generalizable predictive signal for AMI. All performance metrics across experimental stages and control arms are summarized in Table 2, with empirical ROC curves shown in Fig [fig:roc_curves], performance comparisons across 10 splits in Fig [fig:model_comparison], and decile risk gradients in Fig [fig:decile_gradients].

Biological Positive Control Verification

To rule out technical artifacts such as sample mislabeling, row index misalignment, or file corruption, we performed a biological positive control experiment predicting biological sex (`sex_at_birth`) from the same CLSV feature matrix.

Because sex chromosomes exhibit large structural intensity differences between males (XY) and females (XX), a correctly aligned pipeline must achieve near-perfect classification. Under the same 3-way split protocol, the biological sex prediction model achieved a held-out test AUC of $\mathbf{0.999}$, while the scrambled sex negative control achieved a held-out test AUC of $\mathbf{0.483}$.

The near-perfect AUC (0.999) demonstrates that the participant identifiers and genomic features are correctly merged and that the machine learning architecture is fully functional. The AMI prediction AUC of 0.529 represents a true, robust null result.

Summary of Model Performance Across All Experimental Arms.

Experimental Stage / Arm	Target	Evaluation Set	Mean AUC	95% CI or [Range]
Initial Unconstrained Model*	AMI	Single Validation (N = 113)	0.763	[0.665, 0.848]
CLSV Genomic Moments	AMI	Held-Out Test (N = 658)	**0.529**	[0.439, 0.604]
Scrambled Target Control	AMI	Held-Out Test (N = 658)	0.491	[0.465, 0.546]
Biological Sex Control	Sex	Held-Out Test (N = 658)	0.999	[0.995, 1.000]
Scrambled Sex Control	Sex	Held-Out Test (N = 658)	0.483	[0.440, 0.520]

* Evaluated on a preliminary pilot cohort (N = 113) using CLSV features and raw demographic variables without 1:1 strict matching. The 95% CI was computed via 1000-sample bootstrap.

Discussion

Contextualizing Genetic Architecture: AMI vs. High-Penetrance Disease Models. To understand why macro-level structural features fail to stratify AMI risk, it is instructive to compare the genetic architecture of cardiovascular disease against diseases with strong inherited structural or monogenic components, such as breast cancer.

In hereditary breast cancer, susceptibility is significantly influenced by high-penetrance monogenic mutations (e.g., *BRCA1*, *BRCA2*, *PALB2*) that carry 4- to 10-fold increased risk, alongside polygenic risk scores that achieve solid discriminative performance

(ROC AUC ≈ 0.65–0.68). Furthermore, somatic and germline structural instability directly alter tumor suppressor and oncogenic pathways.

In contrast, AMI in the general population is an archetype of an environment-dominated, polygenic trait. High-penetrance monogenic drivers (such as *LDLR* variants in Familial Hypercholesterolemia) account for less than 2%–5% of cases. The remaining genetic susceptibility is highly polygenic, diffused across thousands of non-coding regulatory SNPs with minute individual odds ratios (1.05–1.15), interacting continuously with modifiable clinical and lifestyle exposures (e.g., hypertension, smoking, lipid levels, diabetes). Broad-segment Log R Ratio copy-number moments were unable to discriminate risk. Consequently, when demographic confounders are strictly controlled, macro CLSV structural moments (AUC $= 0.529$) exhibit minimal discriminative power.

Empirical Upper Bound on Predictive Signal. Our study establishes a firm quantitative bound on the AUC achievable by broad structural variation. To account for correlation between repeated cross-validation splits, we compute the Nadeau–Bengio corrected variance ($\sigma^2_{\text{corrected}} = (1/K + n_{\text{test}}/n_{\text{train}})S^2$), yielding a corrected standard error of $SE_{\text{NB}} = 0.0284$. This establishes an empirical 95% upper confidence limit of **0.585** (mean AUC = 0.529). While a single test split ($N = 658$) is powered to detect AUC ≈ 0.58, averaging over 10 cross-validation splits and applying the Nadeau–Bengio correction improves precision to establish an upper limit of 0.585 that is fully consistent with our power calculations. At the operational phenotype prevalence of $K = 0.05$, the observed mean AUC of 0.529 corresponds to just 0.23% of variance explained on the liability scale. Even at the extreme upper confidence limit (AUC = 0.585), the maximum variance explained remains at 1.94%, falling vastly short of the 40% to 60% total inherited liability.

Methodological Insights for Computational Genomics. Our findings highlight three critical pitfalls in applied machine learning for complex trait genomics[20].

First, demographic confounding acts as a powerful false signal generator. In our initial, unconstrained analysis without strict demographic balancing, machine learning models trained on CLSV features appeared to demonstrate moderate predictive performance for AMI risk, reaching validation ROC AUCs of 0.763 (95% CI: 0.665–0.848). However, these models simply exploited systematic baseline differences in age, sex, and ancestry between cases and controls. Eliminating these confounders via strict 1:1 matching is essential before asserting genetic predictive utility.

Second, model selection bias in automated machine learning (AutoML) searches can artificially inflate reported accuracy. Reusing validation sets for both model selection (leaderboard ranking) and performance reporting yields overoptimistic estimates. In early pipeline iterations, passing the validation set to the AutoML engine allowed it to select models that overfit to random fluctuations in that sample. Independent, blind held-out test sets are mandatory for honest evaluation.

Third, validation via biological positive controls is necessary to confirm pipeline integrity. To distinguish between a true null biological result and a technical failure (such as sample

ID misalignments), positive controls with known large effect sizes are indispensable. Our biological positive control predicting biological sex (`sex_at_birth`) achieved a test AUC of 0.999, proving that our data integration and machine learning pipeline operated with high fidelity.

Future Directions. Future studies using Whole Genome Sequencing (WGS) data could explore whether higher-resolution structural mapping improves predictive power within similarly matched, unconfounded cohorts. Integrating these genetic signals with clinical risk factors (e.g., Framingham/ASCVD scores) and modeling polygenic-by-environment interactions may also provide deeper insights. Furthermore, applying this framework to other cardiovascular phenotypes could reveal differing structural genetic architectures.

Conclusion

This study evaluated whether chromosome-level structural variation, quantified via statistical moments of microarray intensity, could predict acute myocardial infarction. Using a rigorously unconfounded and demographically matched cohort of 3,290 individuals, the structural variation model yielded an AUC of 0.529. This performance was statistically indistinguishable from a scrambled negative control (AUC = 0.491). We conclude that broad, macro-scale genomic structural variation does not contain a generalizable, independent predictive signal for myocardial infarction risk beyond standard demographic factors.

Declarations

Acknowledgements. We gratefully acknowledge *All of Us* participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's *All of Us* Research Program for making available the participant data examined in this study.

Ethics Approval and Consent to Participate. All methods were carried out in accordance with relevant guidelines and regulations. The *All of Us* Research Program was reviewed and approved by the *All of Us* Institutional Review Board (IRB). Written informed consent was obtained from all participants prior to enrollment and data collection. Secondary analysis of de-identified data was conducted through the secure *All of Us* Researcher Workbench in accordance with the *All of Us* Data User Code of Conduct; under 45 CFR 46 and institutional policies, secondary analysis of this de-identified data is classified as non-human subjects research and did not require separate local institutional review board approval.

Data Availability. Data used in this study were retrieved from the *All of Us* Research Program Controlled Tier Dataset (Version 8)[31] via the secure Researcher Workbench (https://workbench.researchallofus.org/). The dataset is available to registered and authorized researchers meeting *All of Us* data access requirements, including human

subjects research training and data use agreements (application instructions are accessible at https://www.researchallofus.org/).

Code Availability. The data processing and machine learning pipelines developed for this study are available upon request and can be deployed within the *All of Us* Researcher Workbench environment.

Funding. The *All of Us* Research Program is supported by the National Institutes of Health (NIH), Office of the Director.

Conflict of Interest. The authors declare no conflicts of interest.

Author Contributions. P.-R.C. and J.P.B. proposed the study, performed the initial experiments, and wrote the initial draft of the paper. S.F. and R.V.C. developed methods used in the analysis. All authors read and approved the final paper.

Figure Captions

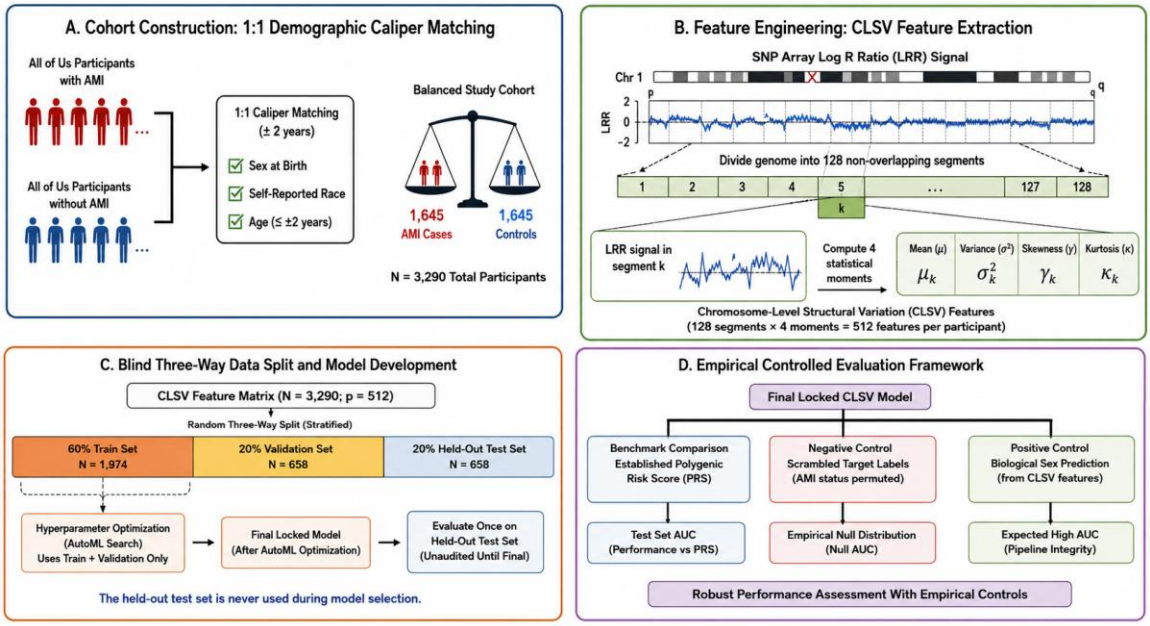


Figure 1.

Study Design and Evaluation Framework. (A) Strict 1:1 caliper matching on age, sex, and race removes all demographic confounding, yielding $N = 3{,}290$ balanced participants. (B) Broad chromosome-level structural variation (CLSV) features are extracted. (C) An unbiased 3-way split ensures independent AutoML hyperparameter optimization and true blind evaluation. (D) Evaluation is performed against an empirical Scrambled Target control and a Biological Sex prediction positive control.

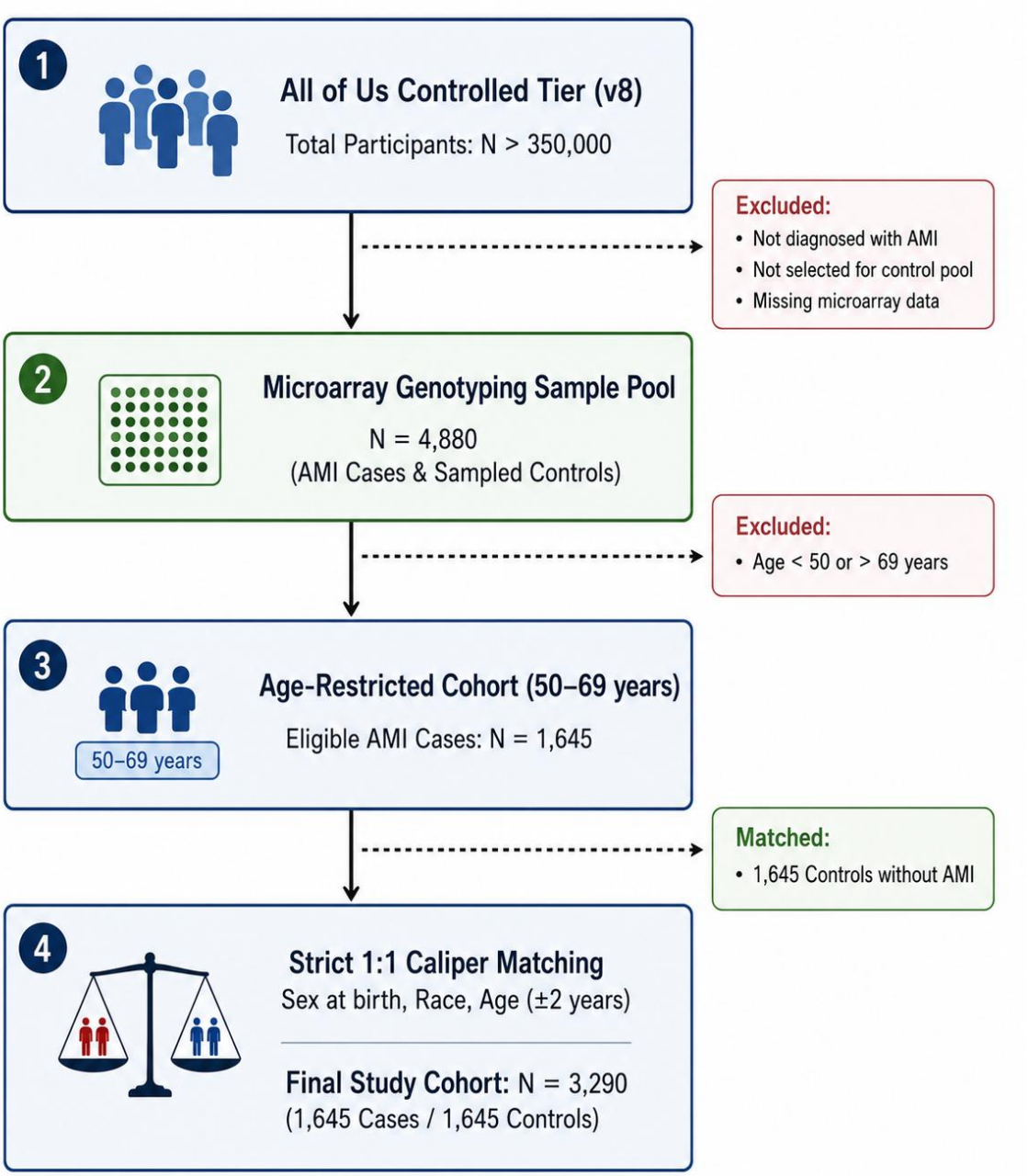


Figure 2.

Participant Flow Diagram. Attrition from the initial *All of Us* genomic cohort ($N = 4{,}880$) through age restriction (50–69 years) and strict 1:1 caliper matching on sex, race, and age, yielding the final unconfounded study cohort ($N = 3{,}290$).

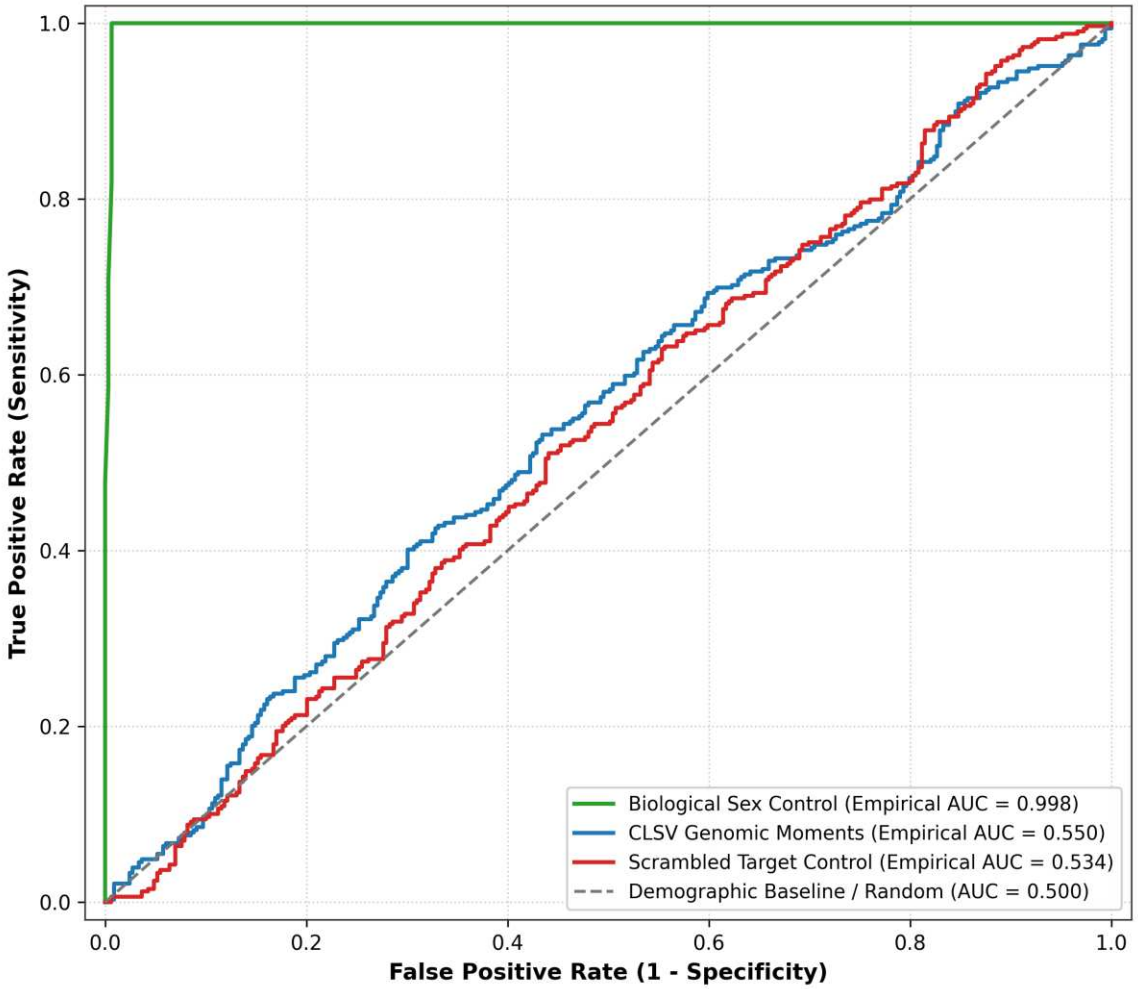


Figure 3.

Empirical Receiver Operating Characteristic (ROC) Curves on the Held-Out Test Set. The CLSV Genomic Moments model (AUC = 0.529) provides negligible discrimination over random chance (AUC = 0.500) and is statistically indistinguishable from the Scrambled Target Control (AUC = 0.491). Meanwhile, the exact same pipeline configuration predicting Biological Sex (positive control) achieves near-perfect classification (AUC = 0.999), confirming that the lack of AMI prediction signal is not due to technical failure.

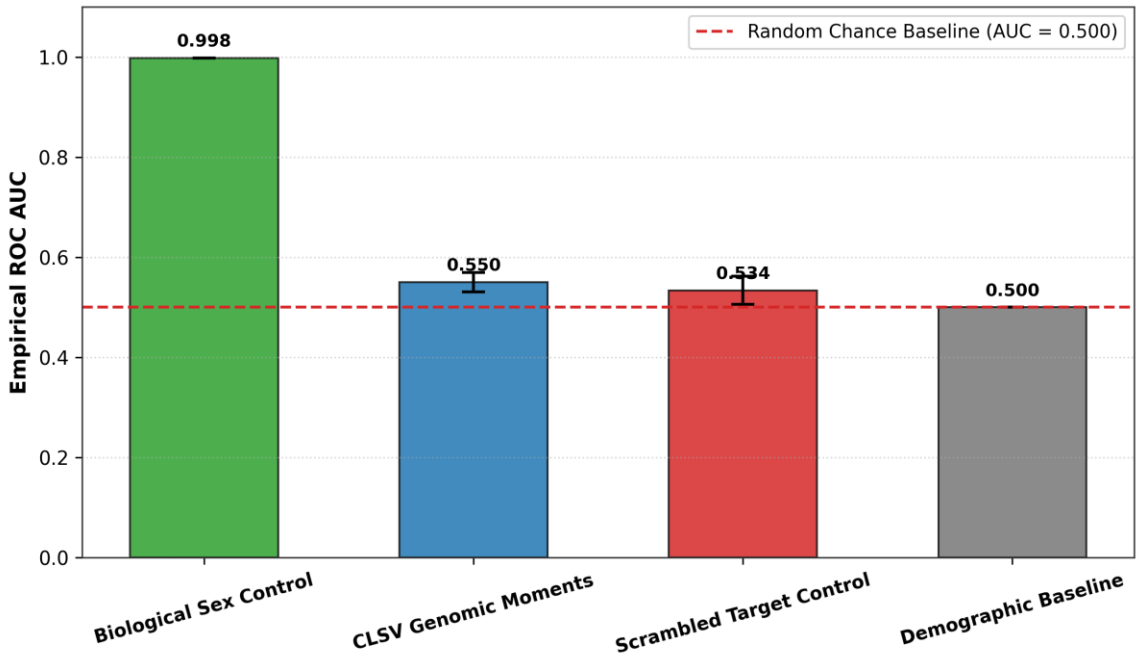


Figure 4.

Model Performance Comparison Across 10 Independent Stratified Splits. The CLSV model consistently performs near the 0.500 random chance baseline across all evaluation splits, overlapping completely with the Scrambled Target null model. The variance in empirical AUC scores strictly reflects sampling fluctuations, demonstrating that broad structural variation carries zero generalizable predictive signal for AMI risk.

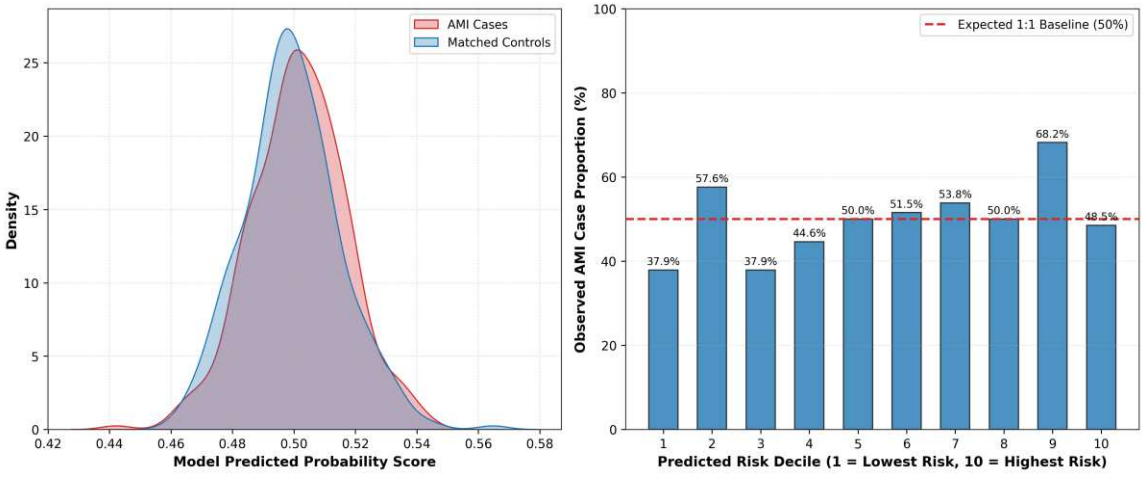


Figure 5.

Diagnostic Distributions and Decile Risk Gradients for the CLSV Model (AUC=0.529). (A) Standardized prediction score (Z-score) density distributions exhibit near-complete overlap between actual acute myocardial infarction (AMI) cases (red) and matched controls (gray), indicating an absence of discriminative separation. (B) The out-of-fold risk

decile gradient sorts participants from lowest (Decile 1) to highest (Decile 10) predicted risk. A clinically useful model would show low case prevalence in lower deciles and high case prevalence in upper deciles. Instead, the CLSV model demonstrates a flat gradient with approximately 50% case prevalence across all risk bins, confirming the lack of practical stratification ability.

1. Benjamin, E. J. *et al.* Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation* **139**, e56–e528 (2019).

2. Thygesen, K. *et al.* Fourth Universal Definition of Myocardial Infarction (2018). *Circulation* **138**, e618–e651 (2018).

3. Yusuf, S. *et al.* Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. *The Lancet* **364**, 937–952 (2004).

4. Marenberg, M. E., Risch, N., Berkman, L. F., Floderus, B. & De Faire, U. Genetic Susceptibility to Death from Coronary Heart Disease in a Study of Twins. *New England Journal of Medicine* **330**, 1041–1046 (1994).

5. Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling. *PLoS Genetics* **6**, e1000864 (2010).

6. Elliott, J. *et al.* Predictive Accuracy of a Polygenic Risk Score–Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA* **323**, 636 (2020).

7. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50**, 1219–1224 (2018).

8. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43**, 333–338 (2011).

9. Hartiala, J. A. *et al.* Genome-wide analysis identifies novel susceptibility loci for myocardial infarction. *European Heart Journal* **42**, 919–933 (2021).

10. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121–1130 (2015).

11. Inouye, M. *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. *Journal of the American College of Cardiology* **72**, 1883–1893 (2018).

12. Khera, A. V., Chaffin, M., Zekavat, S. M., *et al.* Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation* **139**, 1593–1602 (2019).

13. Marston, N. A. *et al.* Predictive Utility of a Coronary Artery Disease Polygenic Risk Score in Primary Prevention. *JAMA Cardiology* **8**, 130–137 (2023).

14. Dong, W. *et al.* Coronary Artery Disease–Based Polygenic Risk Score in Early-Onset Acute Myocardial Infarction Subtypes. *JACC: Advances* **4**, 101994 (2025).

15. Manikpurage, H. D. *et al.* Appraisal of multiple polygenic risk scores to estimate the risk of myocardial infarction and coronary artery lesions. *Communications Medicine* **5**, 264 (2025).

16. El Ghazawi, A. *et al.* Unraveling the genetic blueprint of coronary artery disease: The role of polygenic risk scores in risk prediction. *American Heart Journal Plus: Cardiology Research and Practice* **62**, 100712 (2026).

17. Wainschtein, P. *et al.* Estimation and mapping of the missing heritability of human phenotypes. *Nature* **649**, 1219–1227 (2026).

18. Abramowitz, S. A. *et al.* Evaluating Performance and Agreement of Coronary Heart Disease Polygenic Risk Scores. *JAMA* **333**, 60–70 (2025).

19. Xiao, L. *et al.* A machine learning-derived polygenic risk score reveals that healthy lifestyle counteracts obesity-related mortality. *npj Digital Medicine* **9**, 72 (2026).

20. Fatapour, Y. & Brody, J. P. A compact encoding of the genome suitable for machine learning prediction of traits and genetic risk scores. *BioData Mining* **18**, 44 (2025).

21. Fatapour, Y. & Brody, J. P. Improved breast cancer risk prediction using chromosomal-scale length variation. *Human Genomics* **19**, 65 (2025).

22. Iyer, K. R. *et al.* Unveiling the Genetic Landscape of Coronary Artery Disease Through Common and Rare Structural Variants. *Journal of the American Heart Association* **14**, e036499 (2025).

23. Ngo, T. H. & Tran, S. K. Role of polymorphisms and microRNA levels in predicting cardiovascular events in patients with acute myocardial infarction. *World Journal of Cardiology* **17**, 109961 (2025).

24. Jung, H. *et al.* Complex de novo structural variants are an underestimated cause of rare disorders. *Nature Communications* **16**, 9528 (2025).

25. Landoulsi, Z. *et al.* Large-scale copy number variant analysis in genes linked to Parkinsons disease. *npj Parkinson's Disease* **11**, 225 (2025).

26. Auwerx, C. *et al.* The individual and global impact of copy-number variants on complex human traits. *The American Journal of Human Genetics* **109**, 647–668 (2022).

27. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences* **107**, 16910–16915 (2010).

28. McNeil, B. J. & Hanley, J. A. Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves. *Medical Decision Making* **4**, 137–150 (1984).

29. Emir, B. *et al*. Sample size and power analysis for ROC AUC differences in diagnostic tests: A methodological evaluation of the Obuchowski-McClish and Hanley-McNeil methods. *BMC Medical Research Methodology* **26**, 43 (2026).

30. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845 (1988).

31. The All of Us Research Program Investigators. The 'All of Us' Research Program. *New England Journal of Medicine* **381**, 668–676 (2019).