Exposure to Air Pollutants and Heart Failure Incidence (UK Biobank)

Madison Morales
Brandeis University | RBIF 120 Research Topics in Computational Biology
Dr. Karol Estrada
June 16, 2025

Introduction

Heart failure is a major global health problem and one of the most serious outcomes of cardiovascular disease. Recent studies suggest that long-term exposure to air pollution—especially pollutants like $PM_{2.5}$, NO_2, and NO_x—may contribute to heart failure by triggering inflammation and oxidative stress in the body (Brook et al., 2004; Ghio et al., 2003). These pollutants are known to affect the lungs and heart by increasing reactive oxygen species and disrupting normal cellular processes (Li et al., 2003; Fuertes et al., 2020).

At the same time, researchers have developed polygenic risk scores (PRS) to estimate a person's genetic risk for complex conditions like heart failure. One large study using UK Biobank data found that people with both high genetic risk and high exposure to air pollution had a much greater risk of having a heart attack than those with either risk factor alone (Ma et al., 2024). Another recent study using a dataset of over 50 million individuals emphasized the importance of disentangling genetic heritability, environmental risk, and causal effects of air pollution on cardiovascular disease (Markus et al., 2024). However, fewer studies have explored whether these combined effects also increase the risk of developing heart failure.

This project aims to fill that gap by testing whether long-term exposure to air pollution interacts with genetic risk to influence heart failure incidence. Using data from the UK Biobank (Bycroft et al., 2018), we looked at three pollutants ($PM_{2.5}$, NO_2, NO_x) and tested their effects in people with different levels of polygenic risk for heart failure. These pollutant values came from the ESCAPE Land Use Regression (LUR) model (Field IDs 24003–24008) (UK Biobank, 2021; de Hoogh et al., 2016). The polygenic scores were provided by Dr. Estrada (see Methods).

In addition, we studied three protein biomarkers—interleukin-6 (IL-6), superoxide dismutase 1 (SOD1), and superoxide dismutase 2 (SOD2)—using proteomic data from the Olink 3000 panel (Assarsson et al., 2014). These proteins are linked to inflammation and antioxidant defense and may help explain how pollution leads to disease (Niu et al., 2022; Wang et al., 2019). IL-6, a pro-inflammatory cytokine elevated in response to pollutant exposure, has been associated with cardiovascular risk in both controlled and population-based studies (Thompson et al., 2010).

We hypothesized that: (1) individuals with higher exposure to air pollution and higher genetic risk would have a greater chance of developing heart failure, and (2) the protein biomarkers would show different patterns depending on pollutant levels, reflecting biological stress from the environment.

This research brings together environmental exposure data, genetic information, and proteomics to better understand how air pollution may contribute to heart failure. By studying these connections, we hope to identify biological mechanisms that could help predict or prevent pollution-related cardiovascular disease.

Methods

Study Population and Data Sources

This study used data from the UK Biobank, a large prospective cohort of approximately 500,000 participants aged 40–69 at baseline. Two analytical datasets were created: one including phenotype and air pollution data (n = 439,509), and a subset also containing proteomic data (n = 45,260). Participants missing all six air pollution variables (n = 7,357) were excluded, but individuals missing only some pollutants—particularly PM measures—were retained to preserve sample size. Proteomics missingness ranged from 2.80% to 4.78%, consistent with expectations for high-throughput datasets.

Only complete-case datasets were used in the final analyses to improve model performance and interpretability. Variables with high collinearity (e.g., weight) were removed during model refinement; body fat percentage was retained due to its relative independence from other anthropometric measures.

Environmental Exposure Variables

Air pollution exposure was based on 2010 annual average residential concentrations derived from the ESCAPE Land Use Regression (LUR) model. Pollutants included nitrogen dioxide (NO_2), nitrogen oxides (NO_x), particulate matter less than 2.5 μm ($PM_{2.5}$), $PM_{2.5}$ absorbance (black carbon proxy), PM_{10}, and coarse PM ($PM_{2.5-10}$).

All pollutant variables were examined for distributional properties and found to be right-skewed, which is typical for environmental exposure data. As such, log transformation was applied for normalization, following protocols used in prior environmental epidemiology research (Blackwood, 1995).

Genetic and Proteomic Measures

Polygenic risk scores (PRS) for heart failure were calculated in a previous RBIF120 project “AI Based Predictive Modeling of Heart Failure Outperform Single Polygenic Risk Score Predictive Accuracy Through Addition of Phenotypic and Proteomic Data”. Genome-wide association study summary statistics were downloaded from the Shah et al paper (12) with the GWAS catalog study id of GCST009541 to generate PRS on UkBiobank data using PRSice-2 (https://choishingwan.github.io/PRSice/). The PRS was used to stratify participants into tertiles for certain interaction and stratified analyses.

Proteomic expression data were obtained from the Olink Explore 3000 panel and included three biomarkers: interleukin-6 (IL-6), superoxide dismutase 1 (SOD1), and superoxide dismutase 2 (SOD2). The data were rank-inverse normalized by Olink prior. Expression distributions were centered around 0 with tails truncated at ±4.27, consistent with standard Olink practices.

Missingness in the proteomic data was moderate: 4.78% for SOD1, 2.8% for SOD2, and 3.28% for IL-6. These levels are typical of high-throughput proteomic datasets and were handled using complete-case analysis.

Statistical Analyses

Multiple modeling strategies were used to evaluate the association between air pollution, genetic risk, and heart failure. These included:

- Logistic regression to assess heart failure status across pollution levels and genetic risk
- Linear regression to evaluate associations between pollution levels and protein expression
- Stepwise logistic regression (based on AIC) to select optimal covariates
- Variance Inflation Factor (VIF) analysis to identify and address multicollinearity
- Logistic regression with interaction terms between pollution and PRS
- Tertile synergy analysis with categorical interaction terms
- Stratified logistic regression within PRS tertiles

Covariates included in each model varied depending on data availability and model performance. Commonly adjusted phenotypic variables included age, sex, and body fat percentage.

	Dataset	**Covariates**	**Analysis**	**Reason For Analysis**	**Notable Significant Findings (not including phenotype data like age, sex, BMI, etc.)**
Pollution + Phenotype					
	Pollution + Phenotype	age, sex, BMI, NO_2, $PM_{2.5}$	Logistic Regression	First analysis	$PM_{2.5}$ ($p=0.0194$)
	Pollution + Phenotype	age, sex, BMI, waist circumference, body fat percent, NO_2, NO_x, PM_{10}, $PM_{2.5}$, $PM_{2.5}$absorbance, $PM_{2.5-10}$.	Expanded Logistic Regression	Testing all variables regardless of collinearity	Just phenotypic data
	Pollution + Phenotype	age, sex, waist circumference, body fat percent, NO_2, NO_x, $PM_{2.5}$	Stepwise Logistic Regression	Select optimal set of predictors based on AIC to improve model fit and reduce overfitting.	NO_2, NO_x ($p<0.05$)
	Pollution + Phenotype	age, sex, waist circumference, body fat percent, NO_2, NO_x, $PM_{2.5}$	Variance Inflation Factor (VIF)	Assess multicollinearity and identify correlated predictors that may compromise model stability or interpretability.	NA
	Pollution + Phenotype	age, sex, waist circumference, body fat percent, NO_2, NO_x, $PM_{2.5}$	Logistic Regression	Final reduced model based on stepwise AIC selection, retaining strongest phenotype + pollution variables while balancing collinearity and predictive power.	NO_2 ($p = 0.0345$), NO_x ($p = 0.0476$)

	Pollution + Phenotype	age, sex, body fat percent, NO_2, $PM_{2.5}$, NO_2:age, NO_2:sex, $PM_{2.5}$:age, $PM_{2.5}$:sex	Logistic Regression with Interaction Terms	To explore whether the effect of air pollution on heart failure risk differs by age or sex through interaction terms.	Just phenotypic data
	Pollution + Phenotype	age, sex, waist circumference, body fat percent, NO_2	Logistic Regression	To isolate and evaluate the independent effect of NO_2on heart failure by removing correlated air pollution variables (e.g., NO_x, $PM_{2.5}$), which could introduce multicollinearity.	Just phenotypic data
Pollution + Phenotype + Proteomics					
	Pollution + Phenotype + Proteomics	age, sex, BMI, NO_2, $PM_{2.5}$, SOD1, SOD2, IL-6	Logistic Regression	First analysis	IL-6 ($p<0.001$)
	Pollution + Phenotype + Proteomics	age, sex, BMI, waist circumference, body fat percent, NO_2, NO_x, PM_{10}, $PM_{2.5}$, $PM_{2.5}$absorbance_2010, $PM_{2.5}$_10_2010, SOD1, SOD2, IL-6	Expanded Logistic Regression	Testing all variables regardless of collinearity	IL-6 ($p<0.001$)
	Pollution + Phenotype + Proteomics	same as model_expanded	Variance Inflation Factor (VIF)	Assess multicollinearity and identify correlated predictors that may compromise model stability or interpretability.	High VIF for NO_2(12.2), NO_x (7.7), body fat variables (VIF >5)
	Pollution + Phenotype + Proteomics	age, sex, body fat percent, NO_x, $PM_{2.5}$, IL-6	Stepwise Logistic Regression	Select optimal set of predictors based on AIC to improve model fit and reduce overfitting.	IL-6 ($p<0.001$)
	Pollution + Phenotype + Proteomics	age, sex, body fat percent, NO_x, $PM_{2.5}$, IL-6	Logistic Regression	Final reduced model chosen based on stepwise selection, multicollinearity diagnostics (VIF), statistical significance, and AIC minimization.	IL-6 ($p = 0.000268$)
	Pollution + Phenotype + Proteomics	age, sex, body fat percent, NO_2, $PM_{2.5}$, IL-6, NO_2:IL-6, $PM_{2.5}$:IL-6	Logistic Regression with Interaction Terms	To test whether the inflammatory marker IL-6 modifies the association between air pollution exposure and heart failure incidence.	Just phenotypic data

	Pollution + Phenotype + Proteomics	$PM_{2.5}$, NO_2, NO_x, age, sex, body fat percet	Linear Regression	To test whether exposure to air pollutants ($PM_{2.5}$, NO_2, NO_x) is associated with systemic inflammation, using IL-6 as a biomarker. This evaluates the biological plausibility of pollution-induced cardiovascular risk through inflammatory pathways.	$PM_{2.5}$ (p = 0.000564)
	Pollution + Phenotype + Proteomics	$PM_{2.5}$, NO_2, NO_x, age, sex, body fat percent	Linear Regression (semi-partial R^2 for PM2.5)	To quantify the unique contribution of PM2.5 to IL-6 expression by calculating its semi-partial R^2 through comparison with a reduced model excluding PM2.5.	NA
Pollution + Phenotype + PRS					
	Pollution + Phenotype + PRS	age, sex, waist circumference, body fat percent, NO_2, NO_x, $PM_{2.5}$, PRS	Logistic Regression	Assess whether the inclusion of a polygenic risk score (PRS) improves model performance and helps clarify the relationship between pollution and heart failure.	NO_2 (p < 2e-16), NO_x (p = 1.33e-05), $PM_{2.5}$ (p = 1.75e-06), PRS (p < 2e-16)
	Pollution + Phenotype + PRS	age, sex, body fat percent, NO_2, NO_x, $PM_{2.5}$, PRS, PRS:NO_2, PRS:$PM_{2.5}$, PRS:NO_x	Logistic Regression with Interaction Terms	To test whether the effect of pollution on heart failure risk is modified by genetic susceptibility (gene–environment interaction).	NO_2 (p = 3.77e-15), NO_x (p = 0.001085), $PM_{2.5}$(p = 0.000221), PRS (p = 0.002360), NO_2 × PRS (p = 1.39e-14), $PM_{2.5}$ × PRS (p = 0.000312), NO_x × PRS (p = 0.001433)
	Pollution + Phenotype + PRS	age, sex, body fat percent, NO_2, PRS, PRS:NO_2	Logistic Regression with Interaction Terms	To isolate the interaction between NO_2 and polygenic risk score (PRS) to test for gene–environment effects.	NO_2 (p < 2e-16), PRS (p < 2e-16), NO_2 × PRS interaction (p < 2e-16)
	Pollution + Phenotype + PRS	age, sex, body fat percent, $PM_{2.5}$, PRS, $PM_{2.5}$:PRS	Logistic Regression with Interaction Terms	To test for gene–environment interaction between $PM_{2.5}$ and genetic risk (PRS) in predicting heart failure	$PM_{2.5}$ (p = 8.08e-09), PRS (p < 2e-16), $PM_{2.5}$ × PRS interaction (p = 1.02e-08)

	Pollution + Phenotype + PRS	age, sex, body fat percent, NO_x, PRS, NO_x:PRS	Logistic Regression with Interaction Terms	To test whether polygenic risk modifies the effect of NO_x on heart failure incidence	NO_x ($p = 1.89e-13$), PRS ($p < 2e-16$), NO_x × PRS interaction ($p = 2.46e-13$)
	Pollution + Phenotype + PRS	PRS tertile, NO_2 tertile, age, sex, body fat percent, and all interaction terms between PRS and NO_2 tertiles	Logistic regression with categorical interaction terms (tertile synergy analysis)	To test for synergistic effects between polygenic risk and NO_2 exposure on heart failure risk by stratifying both variables into tertiles	Medium PRS ($p = 0.0336$), High PRS ($p < 2e-16$)
	Pollution + Phenotype + PRS	PRS tertile, NO_x tertile, age, sex, body fat percent, and all interaction terms between PRS and NO_x tertiles	Logistic regression with categorical interaction terms (tertile synergy analysis)	To test whether the impact of NO_x exposure on heart failure risk differs across levels of genetic risk (PRS)	PRS (High) ($p < 2e-16$)
	Pollution + Phenotype + PRS	PRS tertile, $PM_{2.5}$ tertile, age, sex, body fat percent, and all interaction terms between PRS and $PM_{2.5}$ tertiles	Logistic regression with categorical interaction terms (tertile synergy analysis)	To assess whether the risk associated with $PM_{2.5}$exposure differs across levels of genetic risk (PRS)	PRS (High) ($p < 2e-16$)
	Pollution + Phenotype + PRS	NO_2 tertile (Low [reference], Medium, High), age, sex, body fat percent	Stratified logistic regression within PRS tertiles	To assess how the effect of NO_2 exposure on heart failure risk varies across different levels of genetic risk (PRS), by estimating associations within each genetic stratum using low NO_2 as the reference group	Just phenotypic data
	Pollution + Phenotype + PRS	NO_2 tertile (Low [reference], Medium, High), age, sex, body fat percent	Stratified logistic regression within PRS tertiles	To assess how the effect of NO_2 exposure on heart failure risk varies across different levels of genetic risk (PRS), by estimating associations within each genetic stratum using low NO_2 as the reference group	Just phenotypic data

	Pollution + Phenotype + PRS	NO_2 tertile (Low [reference], Medium, High), age, sex, body fat percent	Stratified logistic regression within PRS tertiles	To assess how the effect of NO_2 exposure on heart failure risk varies across different levels of genetic risk (PRS), by estimating associations within each genetic stratum using low NO_2 as the reference group	Just phenotypic data
Pollution + Phenotype + Proteomics + PRS					
	Pollution + Phenotype + Proteomics + PRS	age, sex, body fat percent, NO_2, $PM_{2.5}$, PRS, IL-6, SOD1, SOD2	Logistic Regression	Full model to examine combined effects of demographics, pollution, genetic risk (PRS), and inflammatory/oxidative stress markers on heart failure risk	NO_2 (p = 0.0259), PRS (p < 2e-16), IL-6 (p = 9.10e-05)
	Pollution + Phenotype + Proteomics + PRS	age, sex, body fat percent, NO_2, $PM_{2.5}$, PRS, IL-6, PRS × IL-6, IL-6 × NO_2, IL-6 × $PM_{2.5}$	Logistic Regression	To test whether IL-6 modifies the relationship between air pollution or genetic risk and heart failure risk.	PRS (p < 2e-16)
	Pollution + Phenotype + Proteomics + PRS	age, sex, body fat percent, NO_2, NO_x, $PM_{2.5}$, PRS	Logistic Regression	To evaluate the significance of air pollution variables in a smaller subset with proteomics data, excluding proteins from the model.	NO_2 (p = 0.00231), NO_x (p = 0.02436), PRS (p < 2e-16)
Pollution + Phenotype + Proteomics + PRS + PCA					
	Pollution + Phenotype + Proteomics + PRS + PCA	age, sex, waist circumference, body fat percent, NO_2, NO_x, $PM_{2.5}$, PRS, PC1–PC10	Logistic Regression	To control for population stratification and ancestry-related confounding using principal component (PC) scores	PRS (p < 2e-16), PC1 (p = 7.50e-09), PC4 (p = 5.82e-12)
Pollution + Phenotype + PRS + PCA					
	Pollution + Phenotype + PRS + PCA	age, sex, body fat percent, NO_2, NO_x, $PM_{2.5}$, PRS, IL-6, SOD1, SOD2, PC1–PC10	Logistic Regression	To assess whether pollution, PRS, and proteomic biomarkers remain predictive after adjusting for genetic ancestry	PRS (p < 2e-16), IL-6 (p = 0.000454), PC1 (p = 1.43e-08), PC4 (p = 9.69e-12)

Table 1. Summary of dataset-specific models, covariates, model types, and key findings across all analyses conducted in this study.

All statistical analyses were conducted using R, and the following packages were used: ggplot2, knitr, tidyr, and broom.

To evaluate whether associations between air pollution and heart failure were influenced by unmeasured genetic ancestry, we included the first ten genetic principal components (PCs) as covariates in select models. These PCs were calculated by UK Biobank to represent population structure and were used to reduce confounding due to population stratification. Models adjusted for PCs were compared to unadjusted models to assess the robustness of pollutant and PRS effects.

Results

Six datasets were used in this analysis:

1. A phenotype + pollution dataset (n = 439,509),
2. A phenotype + pollution + PRS dataset (n = 439,509),
3. A phenotype + pollution + proteomics dataset (n = 45,260),
4. A phenotype + pollution + PRS + proteomics dataset (n = 45,260),
5. A phenotype + pollution + PRS + PCs dataset (n = 439,509),
6. A phenotype + pollution + PRS + proteomics Pcs dataset (n = 45,260).

Pollutant Associations with Heart Failure

Depending on the model specification, several pollutants were significantly associated with heart failure incidence. In models using the full phenotype and pollution dataset without adjustment for population structure, $PM_{2.5}$ and NO_x were generally positively associated with heart failure, while NO_2 was inversely associated..

- $PM_{2.5}$ was positively associated with heart failure ($p = 0.0194$).
- NO_2 and NO_x showed marginal significance ($p \approx 0.03–0.05$).

When PRS was added to the model, pollutant associations became stronger and more significant. Results from this model are shown in Figure 1:

- $PM_{2.5}$ tended to show a positive association with heart failure (OR = 1.31, 95% CI: 1.17–1.46, $p = 1.75e\text{-}06$).
- NO_2 often showed a negative association (OR = 0.90, 95% CI: 0.88–0.92, $p < 0.001$).
- NO_x generally showed a modest positive association OR = 1.02, 95% CI: 1.01–1.03, $p = 1.33e\text{-}05$).
- Polygenic risk score (PRS) was consistently and strongly associated with disease risk ($p < .001$), highlighting a substantial genetic contribution.

(See *Adjustment for Population Structure* section for full results).

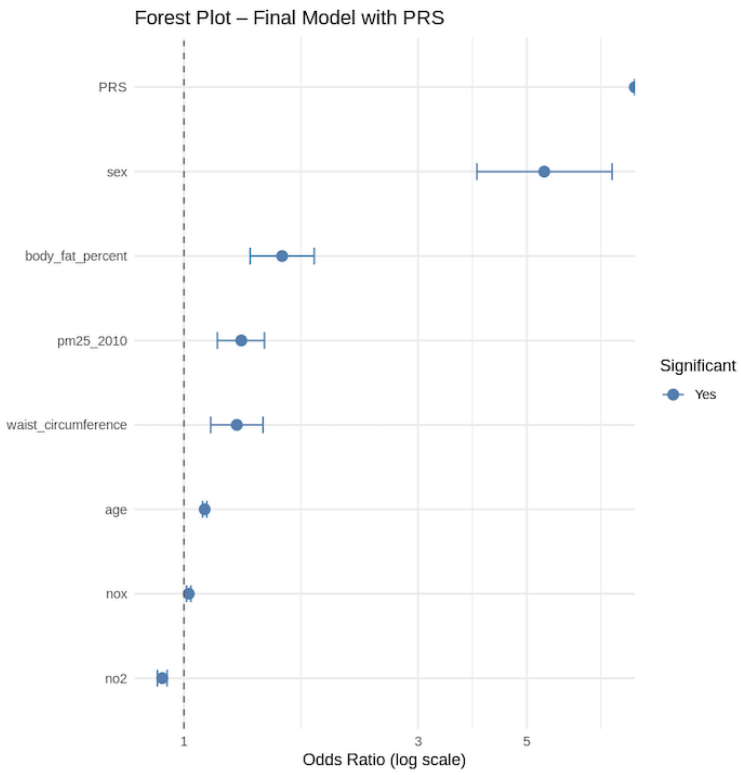

Figure 1. Forest plot showing odds ratios from model including PRS, pollution, and phenotype. PRS, sex, body fat percentage, age, $PM_{2.5}$, NO_x, and NO_2 are significant predictors.

Pollution–Genetic Interactions
To assess whether genetic risk modifies the association between air pollution and heart failure, formal interaction terms between pollutants and PRS were included in multiple logistic regression models. All interaction terms were statistically significant.

- NO_2 × PRS (p= 1.39e-14),
- NO_x × PRS (p = 0.0014),
- $PM_{2.5}$ × PRS (p = 0.00031).

I created additional single interaction models by testing pollutants individually. These models also reached significance for each pollutant.

- NO_2 × PRS (p < 2e-16),
- NO_x × PRS (p = 2.46e-13),
- $PM_{2.5}$ × PRS (p = 1.02e-08).

These interactions were not tested in PCA-adjusted models.

Stratified Models by PRS Tertile
Logistic models were stratified by PRS tertile to further explore gene–environment effects. Across all pollutants, participants in the High PRS group consistently had the highest odds of heart failure, regardless of pollution level. Results from these models are visualized in Figures 2–4.

- NO_2: High PRS × Low NO_2 (OR = 34.19, 95% CI: 29.87–39.15)
- NO_x: High PRS × Medium NO_x (OR = 30.98, 95% CI: 27.15–35.34)
- $PM_{2.5}$: High PRS × Medium $PM_{2.5}$ (OR = 26.82, 95% CI: 23.46–30.66).

In contrast, individuals in the Low PRS group had lower and more variable odds ratios depending on pollution level. For example, Low PRS × Medium NO_2 had an OR of 0.75 (95% CI: 0.37–1.50). These patterns support the strong and consistent role of genetic predisposition, with more variable contributions from pollution exposure.

These full results are summarized in Figure 5, which presents a matrix of corrected odds ratios across PRS and pollution tertiles.

Principal components (PCs) were not included in these stratified models. Adjusting for genetic population structure could potentially impact the observed associations, particularly if PRS and pollution exposure vary across ancestry groups.

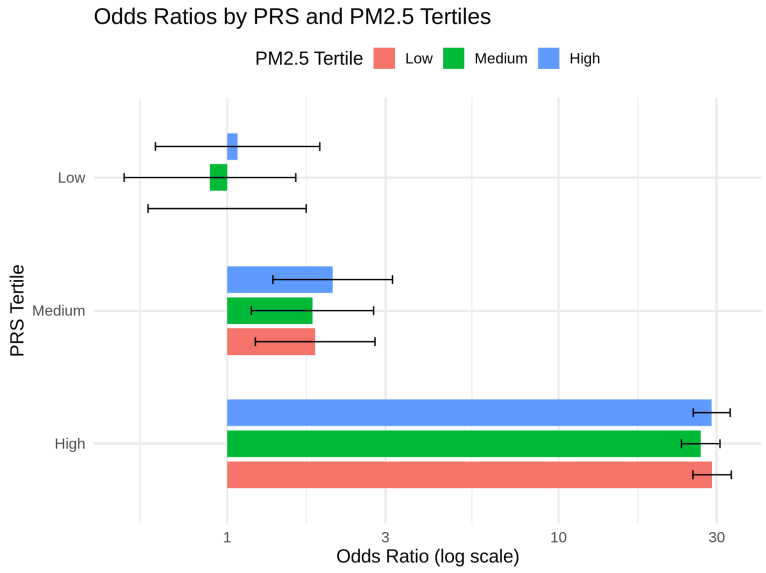

Figure 2. Odds ratios by PRS tertile for $PM_{2.5}$.

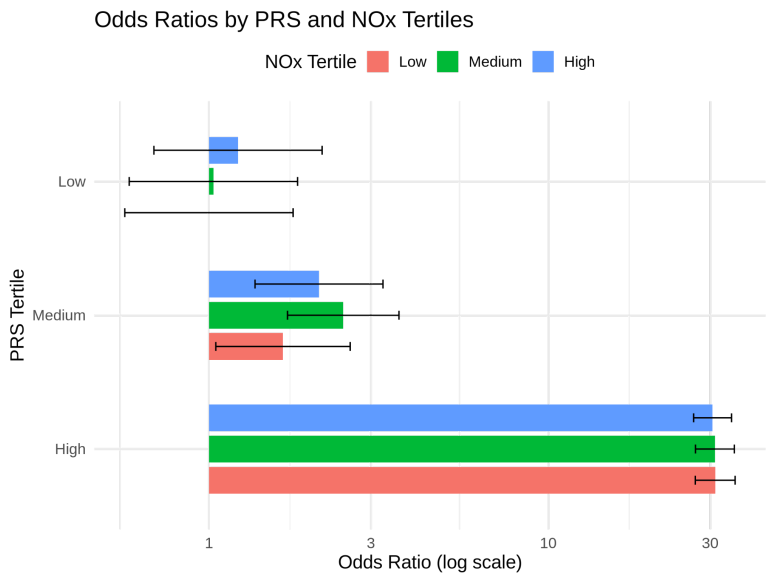

Figure 3. Odds ratios by PRS tertile for NO_x.

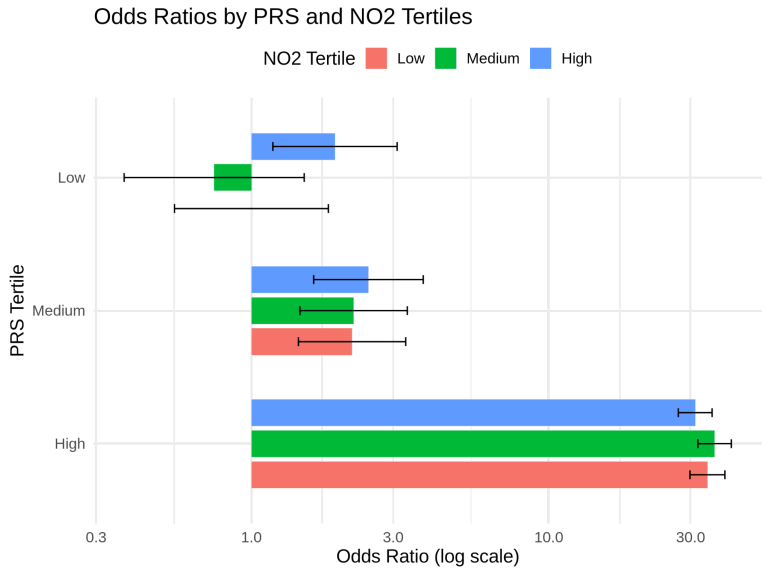

Figure 4. Odds ratios by PRS tertile for NO_2.

Pollutant	PRS_tertile	Pollutant_Tertile	OR	lower_ci	upper_ci
<chr>	<fct>	<fct>	<dbl>	<dbl>	<dbl>
NO2	Low	Low	1.0000000	0.5510563	1.814697
NO2	Medium	Low	2.1802917	1.4389552	3.303558
NO2	High	Low	34.1941413	29.8684798	39.146261
NO2	Low	Medium	0.7493359	0.3731110	1.504926
NO2	Medium	Medium	2.2069377	1.4563501	3.344370
NO2	High	Medium	36.1391090	31.7462413	41.139837
NO2	Low	High	1.9105313	1.1807130	3.091462
NO2	Medium	High	2.4757370	1.6196376	3.784349
NO2	High	High	31.1016395	27.2860011	35.450852
NOx	Low	Low	1.0000000	0.5649519	1.770062
NOx	Medium	Low	1.6527380	1.0476426	2.607323
NOx	High	Low	31.0404296	27.1183406	35.529765
NOx	Low	Medium	1.0307130	0.5822016	1.824745
NOx	Medium	Medium	2.4876445	1.7044095	3.630803
NOx	High	Medium	30.9785738	27.1522753	35.344074
NOx	Low	High	1.2180773	0.6883513	2.155458
NOx	Medium	High	2.1102692	1.3671608	3.257288
NOx	High	High	30.4933748	26.8122750	34.679859
PM2.5	Low	Low	1.0000000	0.5774816	1.731657
PM2.5	Medium	Low	1.8435276	1.2166586	2.793384
PM2.5	High	Low	29.0347564	25.4108769	33.175442
PM2.5	Low	Medium	0.8876509	0.4891069	1.610945
PM2.5	Medium	Medium	1.8087579	1.1828664	2.765828
PM2.5	High	Medium	26.8206114	23.4583714	30.664754
PM2.5	Low	High	1.0754291	0.6076701	1.903249
PM2.5	Medium	High	2.0818630	1.3742206	3.153899
PM2.5	High	High	28.9539245	25.4690282	32.915655

Figure 5. Corrected odds ratios and 95% confidence intervals for heart failure risk across combinations of PRS tertiles (Low, Medium, High) and air pollution exposure tertiles (Low, Medium, High) for NO_2, NO_x, and $PM_{2.5}$. Odds ratios are relative to the Low PRS × Low pollutant exposure group for each pollutant.

Focused Analysis of Low NO_2 Exposure by PRS Tertile

To further investigate the inverse association seen between NO_2 and heart failure in earlier models, a targeted analysis was conducted using only the low NO_2 exposure tertile, stratified by PRS level. This analysis is visualized in Figures 6 and 7.

In the Low PRS group, high NO_2 exposure had a positive association with heart failure (OR = 1.82), though not statistically significant. In the Medium PRS group, NO_2 had minimal effect (ORs = 1.01 and 1.14 for medium and high exposure tertiles, respectively). In the High PRS

group, the trend reversed: medium NO_2 exposure slightly increased risk (OR = 1.06) while high NO_2 showed a negative association (OR = 0.91), again nonsignificant.

These findings suggest NO_2 may have a stronger effect in individuals with lower genetic risk, a trend that diminishes or even reverses at higher PRS levels. Although these interactions were not statistically significant—potentially due to smaller sample sizes—they offer an interesting hypothesis for future study.

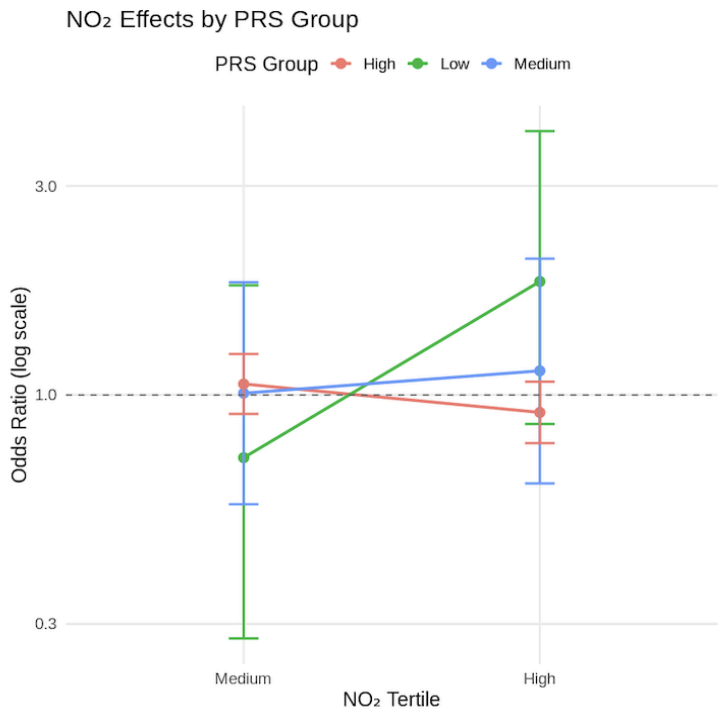

Figure 6. Low NO_2 exposure effect across PRS groups (line plot).

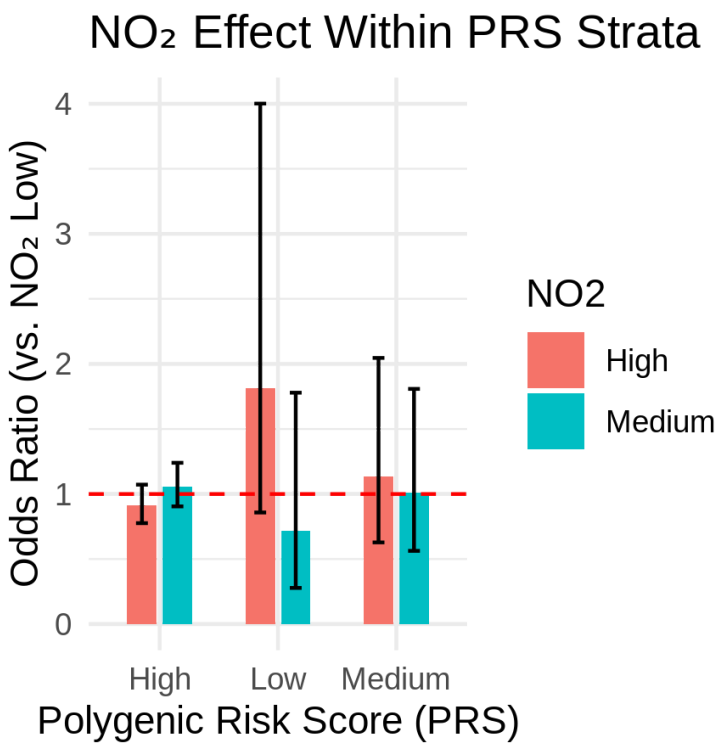

Figure 7. Low NO_2 exposure effect across PRS groups (bar plot).

Proteomics Associations with Pollutants
Logistic regression models were used to evaluate associations between air pollution exposures ($PM_{2.5}$, NO_2, NO_x) and IL-6 expression, adjusted for age, sex, and body fat percentage.

- $PM_{2.5}$ was significantly associated with IL-6 (p = 0.000564)
  - Semi-partial R^2 = 0.00026 (explaining 0.026% of variance in IL-6 expression)
- NO_2 was not statistically significant (p = 0.122)
- NO_x was not statistically significant (p = 0.507)

In logistic regression models using the pollution + phenotype + proteomics dataset:

- IL-6 remained a statistically significant predictor in all model variations
  - Final reduced model: p = 0.000268
- $PM_{2.5}$ was not statistically significant (p = 0.0817)
- NO_x was not statistically significant (p = 0.1003)
- SOD1 and SOD2 were not significant in any proteomics model and were excluded

In the model including pollution, phenotype, proteomics, PRS, and the first ten genetic principal components (PCs):

- IL-6 remained statistically significant (p = 0.000454)

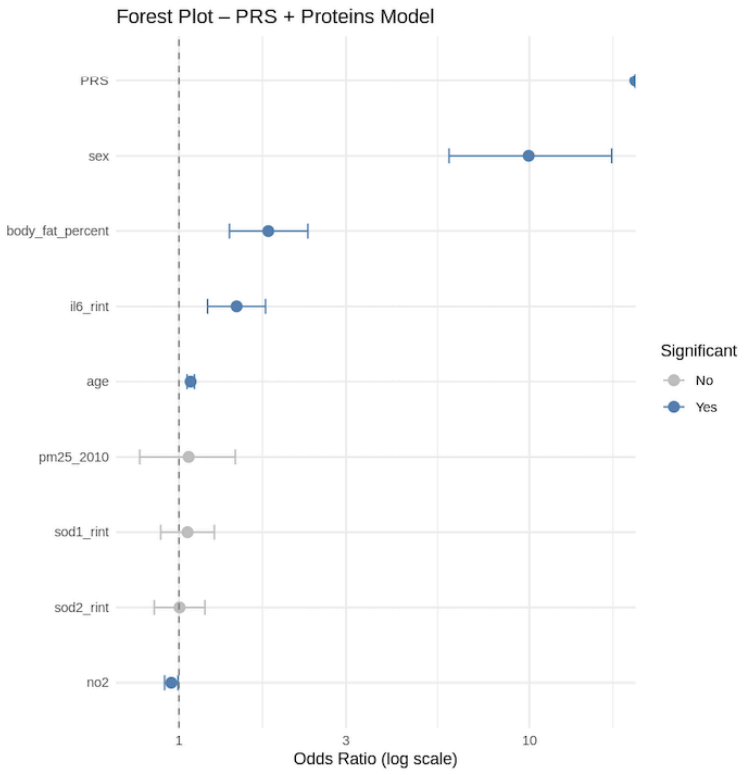

Figure 8. Forest plot for the model including pollution, proteomics, and phenotypic covariates. IL-6 is a significant predictor; pollution variables are attenuated.

Adjustment for Population Structure
To determine whether observed pollution effects could be attributed to population stratification (i.e., ancestry-related confounding), two logistic regression models were run, each incorporating the first ten principal components (PCs) of genetic variation.

Model 1: Included pollution, PRS, and PCs.

- NO_2 (p = 0.356)
- NO_x (p = 0.742)
- $PM_{2.5}$ (p = 0.101)
    - None of the pollutants were statistically significant after adjusting for ancestry.
- PRS remained highly significant (p < 2e-16).
- PRS × NO_2 interaction was not significant (p = 0.176)

Model 2: Included pollution, PRS, proteomics (IL-6, SOD1, SOD2), and PCs.

- NO_2 (p = 0.634)
- NO_x (p = 0.666)
- $PM_{2.5}$ (p = 0.412)
    - Pollutant effects were further diminished and remained non-significant.
- IL-6 remained statistically significant (p = 0.000454).
- PRS remained statistically significant (p < 2e-16).

These findings indicate that population stratification may explain some of the pollution–disease associations seen in simpler models. However, PRS and IL-6 consistently emerged as robust predictors of heart failure risk—even after adjusting for ancestry-related structure.

Discussion

This study examined the associations between air pollution, genetic susceptibility, and biomarkers of inflammation in relation to heart failure incidence. Using UK Biobank data, multiple logistic regression models revealed complex patterns of interaction and stratification. Key findings indicated that $PM_{2.5}$, NO_2, and NO_x were significantly associated with heart failure in models without adjustment for population structure. However, these associations were no longer significant after incorporating genetic principal components, suggesting population stratification may confound pollution-related risk estimates.

PRS was consistently and strongly associated with heart failure across all models, reinforcing the genetic underpinnings of cardiovascular disease risk. Significant interaction terms between PRS and each pollutant (NO_2, NO_x, $PM_{2.5}$) highlighted gene–environment synergy, where the impact of air pollution may be amplified or diminished depending on an individual's genetic profile. Stratified analyses confirmed this trend, with markedly elevated odds ratios in the high PRS group regardless of pollution exposure.

Proteomics analyses further emphasized the role of inflammatory pathways. IL-6 was significantly associated with both $PM_{2.5}$ exposure and heart failure risk, even after adjusting for covariates and principal components, while SOD1 and SOD2 showed no significant associations. These results suggest that IL-6 may serve as a key biomarker linking pollution to cardiovascular disease.

Notably, a focused analysis of NO_2 exposure among PRS strata revealed an intriguing pattern: individuals with lower genetic risk appeared more sensitive to NO_2, whereas those with higher PRS demonstrated an inverse trend. Although this interaction was not statistically significant, the pattern warrants further exploration. Finally, while PCA adjustments clarified potential ancestry confounding, they also underscored the need for caution in interpreting pollutant effects without accounting for underlying genetic structure.

Works Cited

Assarsson, E., Lundberg, M., Holmquist, G., Björkesten, J., Bucht Thorsen, S., Ekman, D., … & Fredriksson, S. (2014). Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLOS ONE*, 9(4), e95192. https://doi.org/10.1371/journal.pone.0095192

Blackwood, L. (1995). *Transformations of environmental variables in ecological studies*. Journal of Applied Ecology, 32(2), 515–521. https://doi.org/10.2307/2404437

Brook, R. D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., ... & Tager, I. (2004). Air pollution and cardiovascular disease: A statement for healthcare professionals from the

expert panel on population and prevention science of the American Heart Association. *Circulation, 109*(21), 2655–2671. https://doi.org/10.1161/01.CIR.0000128587.30041.C8

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., … & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

de Hoogh, K., Gulliver, J., Donkelaar, A. V., Martin, R. V., Marshall, J. D., Bechle, M. J., … & Hoek, G. (2016). Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental Research*, 151, 1–10. https://doi.org/10.1016/j.envres.2016.07.005

Fuertes, E., van der Plaat, D. A., & Minelli, C. (2020). Antioxidant genes and susceptibility to air pollution for respiratory and cardiovascular health. *Free Radical Biology and Medicine, 151*, 181–192. https://doi.org/10.1016/j.freeradbiomed.2020.01.181

Ghio, A. J., Stonehuerner, J., Dailey, L. A., Madden, M. C., & McGee, J. K. (2003). Rapid increases in the steady-state concentration of reactive oxygen species in the lungs and heart after particulate air pollution inhalation. *Environmental Health Perspectives, 111*(7), 749–755. https://doi.org/10.1289/ehp.5986

Li, N., Sioutas, C., Cho, A., Schmitz, D., Misra, C., Sempf, J., Wang, M., Oberley, T., Froines, J., & Nel, A. (2003). Ultrafine particulate pollutants induce oxidative stress and mitochondrial damage. *Environmental health perspectives*, *111*(4), 455–460. https://doi.org/10.1289/ehp.6000

Ma, Y., Li, D., Cui, F., Wang, J., Tang, L., Yang, Y., Liu, R., Xie, J., & Tian, Y. (2024). Exposure to air pollutants and myocardial infarction incidence: A UK Biobank study exploring gene–environment interaction. *Environmental Health Perspectives, 132*(10), 107002. https://doi.org/10.1289/EHP14291

Markus, H., McGuire, D., Yang, L., Xu, J., Montgomery, A., Berg, A., Li, Q., Carrel, L., Liu, D., & Jiang, B. (2024). Dissecting genetic heritability, environmental risk, and causal effects of air pollution using a health insurance database of >50 million individuals. *Penn State College of Medicine*.

Niu, R., Cheng, J., Sun, J., Li, F., Fang, H., Lei, R., Shen, Z., Hu, H., & Li, J. (2022). Alveolar type II cell damage and Nrf2–SOD1 pathway downregulation are involved in PM2.5-induced lung injury in rats. *International Journal of Environmental Research and Public Health, 19*(19), 12893. https://doi.org/10.3390/ijerph191912893

Thompson, A. M. S., Zanobetti, A., Silverman, F., Schwartz, J., Coull, B., Urch, B., Speck, M., Brook, J. R., Manno, M., & Gold, D. R. (2010). Baseline repeated measures from controlled human exposure studies: Associations between ambient air pollution exposure and the systemic inflammatory biomarkers IL-6 and fibrinogen. *Environmental Health Perspectives, 118*(1), 120–124. https://doi.org/10.1289/ehp.0900550

UK Biobank. (2021). *Data on long-term air pollution exposure*. Retrieved from https://biobank.ndph.ox.ac.uk/ukb/ukb/docs/EnviroExposEst.pdf

Wang, L., Luo, H., Wu, S., Yan, X., Liu, J., Wu, Y., & Yang, X. (2019). PM2.5 inhibits SOD1 expression by up-regulating microRNA-206 and promotes ROS accumulation and disease progression in asthmatic mice. *Environmental Pollution, 254,* 113043. https://doi.org/10.1016/j.envpol.2019.113043