Social Determinants of Health and Lifestyle Risk Factors Modulate Genetic Susceptibility for Women's Health Outcomes

Lindsay A Guare, Jagyashila Das, PhD, Lannawill Caruth, Shefali Setia-Verma, PhD

Department of Pathology and Laboratory Medicine, University of Pennsylvania
Philadelphia, PA 19104
Emails: lindsay.guare@pennmedicine.upenn.edu, jagyashila.das@pennmedicine.upenn.edu,
lanna.caruth@pennmedicine.upenn.edu, shefali.setiaverma@pennmedicine.upenn.edu

Abstract

Women's health conditions are influenced by both genetic and environmental factors. Understanding these factors individually and their interactions is crucial for implementing preventative, personalized medicine. However, since genetics and environmental exposures, particularly social determinants of health (SDoH), are correlated with race and ancestry, risk models without careful consideration of these measures can exacerbate health disparities. We focused on seven women's health disorders in the All of Us Research Program: breast cancer, cervical cancer, endometriosis, ovarian cancer, preeclampsia, uterine cancer, and uterine fibroids. We computed polygenic risk scores (PRSs) from publicly available weights and tested the effect of the PRSs on their respective phenotypes as well as any effects of genetic risk on age at diagnosis. We next tested the effects of environmental risk factors (BMI, lifestyle measures, and SDoH) on age at diagnosis. Finally, we examined the impact of environmental exposures in modulating genetic risk by stratified logistic regressions for different tertiles of the environment variables, comparing the effect size of the PRS. Of the twelve sets of weights for the seven conditions, nine were significantly and positively associated with their respective phenotypes. None of the PRSs was associated with different ages at diagnoses in the time-to-event analyses. The highest environmental risk group tended to be diagnosed earlier than the low and medium-risk groups. For example, the cases of breast cancer, ovarian cancer, uterine cancer, and uterine fibroids in highest BMI tertile were diagnosed significantly earlier than the low and medium BMI groups, respectively). PRS regression coefficients were often the largest in the highest environment risk groups, showing increased susceptibility to genetic risk. This study's strengths include the diversity of the All of Us study cohort, the consideration of SDoH themes, and the examination of key risk factors and their interrelationships. These elements collectively underscore the importance of integrating genetic and environmental data to develop more precise risk models, enhance personalized medicine, and ultimately reduce health disparities.

Keywords: Polygenic Risk Scores, Social Determinants of Health, Health Disparities, Genetic Risk, Disease Prediction, Women's Health, Breast Cancer, Endometriosis, Ovarian Cancer, Preeclampsia, Uterine Cancer, Uterine Fibroids

1 Introduction

Since the completion of the Human Genome Project in 2003, countless studies have been conducted to associate genetic variants with diseases[1–3]. However, genetic factors accompanied by environmental factors collectively contribute to pathogenesis and progression of diseases. Therefore, quantifying the effects of multimodal risk factors separately and together will help to improve disease risk models. Accurate stratification of individual disease risk is an essential step in the way to reduce the burden of health disparities and implement personalized preventative care.

For many highly heritable diseases, such as coronary artery disease and type 2 diabetes, PRSs are useful for stratifying patients into risk groups based on their genetics. However, in the context of women's health diseases, which have historically been underfunded[4] and understudied[5], the predictive accuracy of PRSs has been inconsistent, especially across diverse populations[6]. Globally, large efforts have been undertaken to build diverse resources to support such studies, including the UK Biobank[7], Finngen[8], BioVU[9], BioBank Japan[10], the Penn Medicine Biobank[11], and a newer resource funded by the NIH, the All of Us (AOU) Research Program[12]. The growth of large genomic datasets has enabled not only the detection of disease-associated genetic variations but also the possibility of using genetic and non-genetic risk factors to predict disease risk before the onset. Numerous studies, like the WISDOM trial[13] focusing on breast cancer and the eMERGE network examining PRS results for 10 disease outcomes[14], are underway to investigate how PRSs can be incorporated into clinical practice.

Environmental risk factors are multi-faceted, including lifestyle measurements as well as social determinants of health (SDoH). Most of these variables are measured through survey participation. Lifestyle aspects, like alcohol use, smoking, and physical activity, have been linked to disease risk for endometriosis[15], breast cancer[16], and uterine fibroids[17], respectively. SDoH are define measurements for social inequities which can impact a person’s health. These include neighborhood disorder, stress, and loneliness. Chronic stress and loneliness have been shown to increase lifetime risk of many serious diseases, like Alzheimer’s[18], cardiovascular disease[19], etc. Additionally, SDoH impact diseases affecting women specifically[20–22]. Interactions between genetic and environmental effects have been studied previously, with respect to both individual genetic variants[23] and PRSs[24]. It has been shown that incorporating PRS with environment measurements such as stress improves model performance for other complex disorders[25]. Therefore, understanding the influence of lifestyle and environmental factors alongside genetic factors is crucial for predicting women's health outcomes.

One important aspect of predictive modeling in personalized medicine is to examine the disease progression, including the onset of diseases. Both categories of risk factors (genetic and environmental) are most often studied in the context of lifetime disease risk. Time-to-event analyses are growing in popularity to evaluate longitudinal risk, utilizing survival analysis methodologies to evaluate the impact of risk factors on disease progression, including the onset of the disease.

The aim of this study is to identify and quantify interactions between genetic risk of women’s health conditions and external variables in a diverse cohort of women within the AOU. We hypothesize that an individuals’ susceptibility to disease risks is not solely dictated by their genetic composition but is greatly influenced by these environmental and social determinants. Understanding how environmental contexts impact the efficacy and clinical utility of PRSs will help to ensure that they are implemented in equitable ways.

2 Methods

2.1 Study Dataset – All of Us Research Program

The All of Us Research Program (AOU) is a dataset supported by the NIH comprised of 409,420 participants with electronic health record (EHR) data, 245,400 of whom have short-read whole genome sequencing (WGS) data. In our study, we included 145,563 of the WGS individuals who were assigned female at birth[26] For study individuals, genetic ancestry was assigned by the AOU data team, who computed genetic similarity with the 1000 genomes reference populations based on genetic principal components.

The EHR data for AOU are stored as billing codes in tables that follow the Observational Medical Outcomes Partnership (OMOP) structure[27]. For our focus on women's health conditions, we selected breast cancer (BC), cervical cancer (CC), endometriosis (Endo), ovarian cancer (OC), preeclampsia (PE), uterine cancer (UC), and uterine fibroids (UF). Each of these diseases has ICD-9 and ICD-10 diagnosis codes (Results, Table 1). Case/control status was determined by the presence of one or more ICD codes for each phenotype.

2.2 Calculating PRSs for women's health outcomes

The PGS Catalog[28] is a public repository of PRS weights that have been published and validated. We browsed the PGS catalog for PRSs for each condition. In cases when more than one PRS was available, we prioritized sets of weights that had been tested on large, multi-ancestry validation cohorts and that have shown promising results based on metrics such as AUROC. The accession numbers for the weights we selected are shown in Figure 1. We computed all 12 scores from the downloaded files in genome build 38 with Plink 2.0's --score function[29]. The scores for each phenotype were then standardized by genetic ancestry group.

2.3 Environmental variables (SDoH and lifestyle measures)

AOU issued several surveys to its participants, including SDoH and Lifestyle questionnaires, combining instruments from other well-studied surveys. To compute continuous scales for neighborhood physical disorder, neighborhood social disorder, stress, and loneliness, we followed procedures as described in Tesfaye et al 2024[30]. The other two survey-derived lifestyle variables were smoking and alcohol use. For smoking, there were seven questions. For the three quantitative questions (ranging from 0-99), we assigned these values: responses of zero (1), then the remaining quartiles (2-5). For the other four smoking questions, we assigned numeric values to the responses: Not At All (1), Some Days (3), Every Day (5). There were three questions pertaining to alcohol use, and we assigned responses numerical values of one to five, with five corresponding to heavier drinking.

We aimed to capture other health measurements using biometrics and wearables data. Per individual, we used median Body Mass Index (BMI) measurement over time. We quantified activity levels using two Fitbit-derived measurements: daily steps (ST) and daily sedentary minutes (SM), as both have been linked to health risks[31,32]. Similarly to BMI, we took the median across each day that had measurements to obtain one value per individual. Once we computed each of the nine continuous environmental factors, we visualized the Pearson correlation between them to examine how they relate to each other and potentially eliminate any that were highly correlated.

2.4 Statistical analyses

2.4.1 Stratified time-to-event analyses for age at diagnosis

For each case of the six phenotypes, we assigned the age of first diagnosis code of a condition as "age at diagnosis". This age variable was used as outcome for time-to-event analyses. Time-to-event analyses were performed in two different contexts: stratified by genetic risk and stratified by environmental variable level. For each phenotype, we looked at three curves defined by the tertiles of the stratifying variable (low/medium/high). Those curves (1 = low, 2 = medium, 3 = high) were fit to survival functions[33] using KaplanMeierFitter from the lifelines Python package[34]. The three survival functions were compared in a pairwise scheme using the log rank test, which results in a chi-squared test statistic.

2.4.2 Quantifying effects of PRSs in environmental contexts

Association testing was performed for each of the twelve PRSs with their corresponding phenotype. The odds ratio (OR) coefficient was estimated using a logistic regression (with an intercept) in which the outcome was the phenotype, the risk score was the independent variable, and age at the time of the EHR data extraction was included as a covariate (Equation 1).

$$Logit(Phenotype) \sim Intercept + PRS + Age \tag{1}$$

For the phenotypes with more than one set of PRS weights (breast cancer, endometriosis, ovarian cancer, and uterine fibroids), we selected the PRS with the largest regression coefficient, resulting in six phenotypes with significant PRS effects (Results, Figure 1).

Next, for each phenotype and environmental risk factor, we divided our study population into nine groups based on environmental variable tertiles (low, medium, high) and genetic risk tertile (low, medium, high). To illustrate the differences in risk levels among various environmental and genetic risk groups, we used the medium/medium subgroup as a reference and computed the odds ratio (and 95% confidence interval) for the phenotype in each of the other eight subgroups, displayed in 3x3 grids for comparison.

Finally, to examine whether the impact of the polygenic risk score (PRS) on disease risk varied across different levels of environmental risk, we conducted stratified regression analyses. By dividing the study population into subgroups based on environmental factors, we assessed how the association between PRS and disease outcomes changed within each subgroup, allowing us to determine if the PRS effect size was influenced by the level of environmental risk. Each environmental variable was divided into tertiles, and then the logistic regression was performed as described previously (Equation 1) for each of the three subgroups. In a similar manner, we tested the effect of each environmental risk factor on the phenotypes, stratified by genetic risk tertile (Equation 2).

$$Logit(Phenotype) \sim Intercept + Environment + Age \tag{2}$$

3 Results

3.1 PRSs for women's health phenotypes

Our study cohort consisted of female AOU participants with short-read WGS (N = 145,563). We assigned case/control phenotypes in AOU using hierarchical diagnosis billing codes, Table 1considering both ICD-9 and ICD-10 codes, as shown in Table 1.

Table 1: The seven women's health phenotypes tested. The root ICD codes used for case definitions, the number of cases in the female AOU WGS dataset, and the mean age at diagnosis (Dx) for those cases.

Phenotype	ICD-9 Code	ICD-10 Code	AOU Cases	Dx Age Mean (std)
Breast Cancer (BC)	174	C50	6,444	58.4 (11.7)
Cervical Cancer (CC)	180	C53	546	51.1 (13.3)
Endometriosis (Endo)	617	N80	4,306	43.5 (11.6)
Ovarian Cancer (OC)	183	C56	815	55.1 (13.2)
Preeclampsia (PE)	642	O14	1,966	30.3 (7.0)
Uterine Cancer (UC)	182	C55	715	59.1 (11.1)
Uterine Fibroids (UF)	218	D25	10,829	48.2 (11.1)

12 sets of weights selected from PGS catalog with reported associations to our phenotypes of interest were selected (Table 2).

Table 2 : The PRSs evaluated along with their reported traits, number of variants, and the percentage of the population reported as European in development/training (dev) and testing set. Those reported as "Unspecified" did not provide ancestry specific population reporting

Score	Reported Trait	Year	Number of Variants	% EUR in Dev	% EUR in Validation
PGS000004	Breast Cancer	2018	313	100	76.4
PGS004611	Breast Cancer	2023	76	58.6	Unspecified
PGS001299	Cervical cancer	2022	24	100	40
PGS003447	Endometriosis	2021	14	98	54.5
PGS002077	Endometriosis	2022	14	100	37.5
PGS001866	Endometriosis	2022	399	100	37.5
PGS002250	Epithelial ovarian cancer	2022	27,240	100	60
PGS003394	Epithelial ovarian cancer	2022	36	100	50
PGS004593	Preeclampsia	2022	1,102,059	Unspecified	100
PGS001795	Uterine cancer	2023	911,692	83.9	100
PGS001032	Uterine fibroids	2022	161	100	40
PGS002263	Uterine fibroids	2022	4,457	100	100

We tested logistic regressions for each of the 12 sets of weights selected from the PGS catalog. The PRS for each phenotype with the most significant positive effect was chosen for downstream analysis (Figure 1).

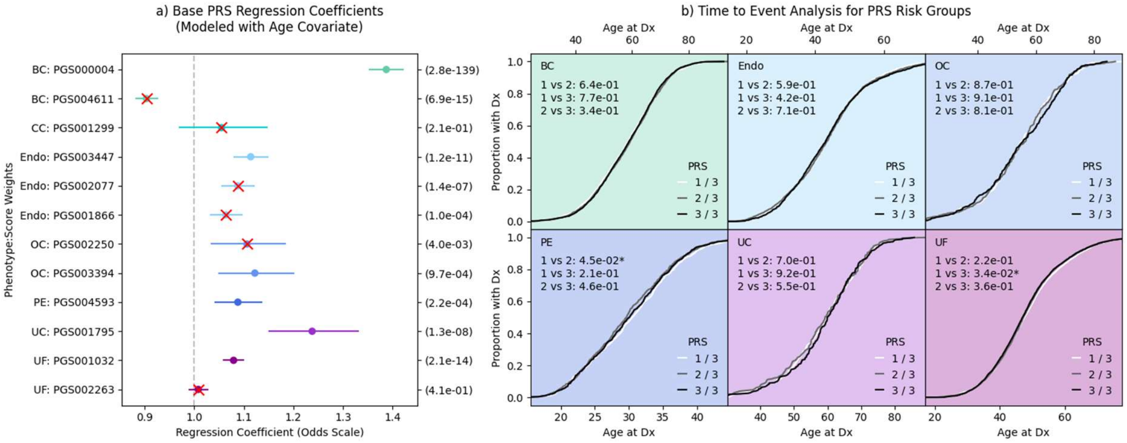


Figure 1: Testing the effects of the PRSs on the women's health outcomes. (a) Coefficients (in odds ratio scale) for logistic regressions based on each PRS. The left axis labels indicate phenotype and PGS Catalog Weights. The right axis labels show the p-value. Scores that were not considered in downstream analyses have a red "X". (b) Time-to event analyses with one curve per PRS risk tertile. Pairwise log rank comparison p values are indicated as text. X-axes above and below each panel are age at diagnosis (Dx). BC: Breast Cancer; UF: Uterine Fibroids; CC: Cervical Cancer; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia.

Based on the logistic regression coefficients for each of the 12 PRSs, we dropped any PRS with odds coefficient <1 (PGS004611 for breast cancer[35]) and any PRS whose p-value for the coefficient was >0.05 (PGS001299 for cervical cancer[36], PGS003394 for ovarian cancer[37], and PGS002263 for uterine fibroids[38]). Since Cervical cancer PRS could not meet these filtering criteria, the phenotype was removed from downstream analysis. In addition, although both PGS002077[39] and PGS001866[39] were significantly associated with endometriosis, only the score that had the strongest effect (PGS003447[40]) was retained.

3.2 Environmental risk factor measurements

The influence of environmental factors, namely, stress level (SL), loneliness level (LL), neighborhood physical disorder (NPD), and neighborhood social disorder (NSD), one biometric measurement (median BMI), two lifestyle scores — alcohol use (AU) and smoking (SK), and two Fitbit measurements — daily steps (ST) and daily sedentary minutes (SM) were tested on susceptibility to genetic risk. We tested these variables for correlation (Figure 2a). Since some measurements were unavailable on all participants, we report the smaller case numbers for each phenotype-measurement combination in Figure 2b.

The most highly correlated variables were NSD and NPD (0.73). Since a higher/greater number of daily steps (ST) is beneficial to health, it was found to be negatively correlated with all other variables except AU. LL was moderately correlated with three other measures, NSD (0.28), NPD (0.21), and SL (0.29). Since some measurements were unavailable for some participants, we report the smaller case numbers for each phenotype-measurement combination. The Fitbit measurements had the fewest participants, so the numbers of cases were small, especially for the rarer phenotypes such as cervical cancer, uterine cancer, ovarian cancer, and preeclampsia. Nearly every participant had BMI measurements, so tests with BMI had the largest sample sizes.

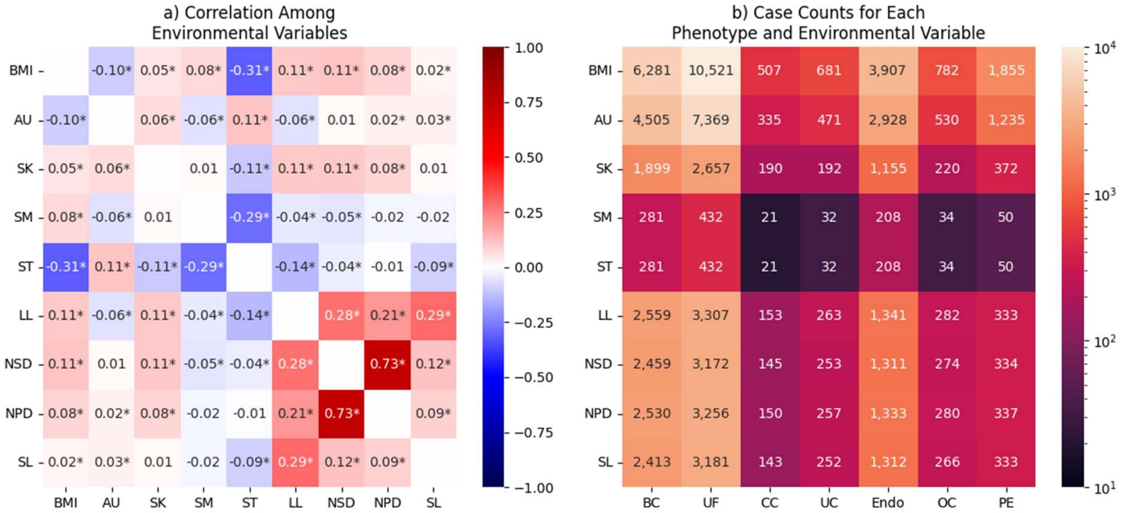
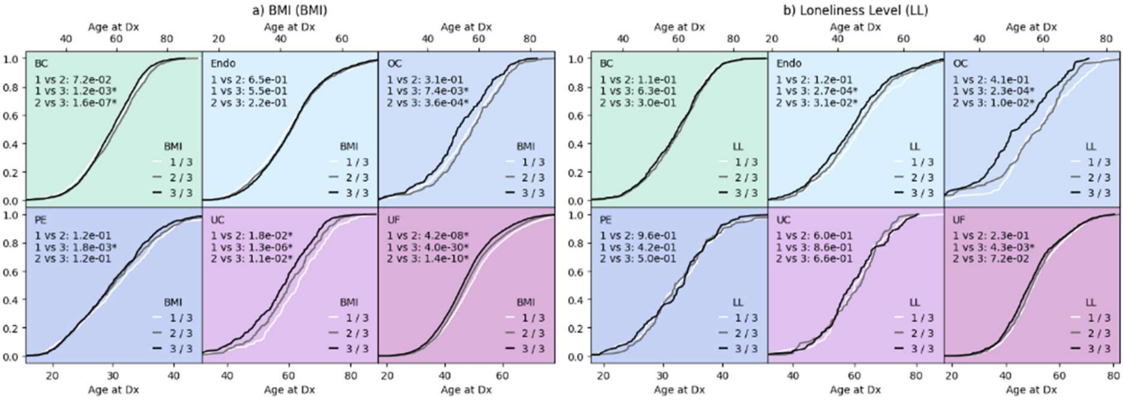


Figure 2: (a) heatmap showing correlation between all nine measurements considered. Correlation values significantly different from zero ($p < 0.05$) are marked with an asterisk. (b) heatmap showing the number of cases for a given phenotype (column) and measurement (row) combination. BC: Breast Cancer; UF: Uterine Fibroids; CC: Cervical Cancer; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia. BMI: Body Mass Index; AU: Alcohol Use; SK: Smoking ; SM: Sedentary Minutes; ST: Steps; LL: Loneliness; NSD: Neighborhood Social Deprivation; NPD: Neighborhood Physical Deprivation and SL: Stress Level.

3.3 Environmental effects on age at diagnosis with time-to-event curves

We estimated the effect of different levels of environmental exposures, categorized into low/medium/high tertiles, on the age at diagnosis for each phenotype. Among the four social determinants of health (SDoH) factors, Neighborhood Social Deprivation (NSD) was removed from the analysis due to its high correlation with Neighborhood Physical Deprivation (NPD), as illustrated previously in Figure 2a. The survival functions, which depict the probability of remaining disease-free over time for each tertile of environmental exposure, are presented in Figure 3. Additionally, the pairwise p-values indicate the statistical significance of the differences between the survival curves for each tertile, highlighting the impact of varying levels of environmental exposures on disease onset.



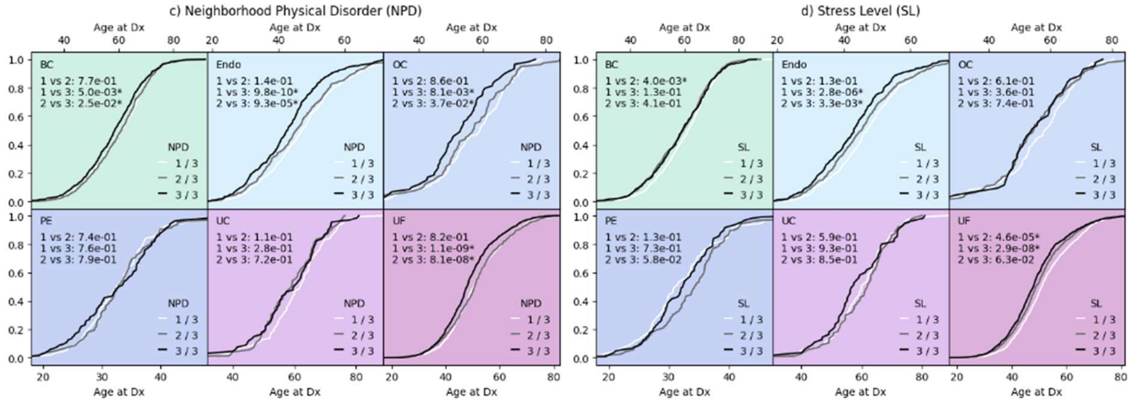


Figure 3: Time-to-event analyses for BMI and the SDoH themes (a - BMI, b - loneliness, c - neighborhood physical disorder, and d - stress). Each panel shows three “survival” curves per phenotype, stratified by the value of the environmental measure where 1 is the lowest tertile and 3 is the highest tertile. The x-axes represent age at diagnosis (Dx). Also indicated in each grid cell are the p-values of pairwise log rank comparisons between those three curves. Any p-values less than 0.05 are annotated with an asterisk. BC: Breast Cancer; UF: Uterine Fibroids; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia.

Of all the environmental risk factors, BMI had the most significant effect on the age at diagnosis. High BMI corresponded to earlier diagnoses of uterine cancer and uterine fibroids (three out of three pairwise comparisons significant), breast cancer and ovarian cancer (two out of three significant), and preeclampsia (P = 1.8 x 10^{-3} comparing first and third tertiles). Those with high LL scores tended to have earlier diagnoses of endometriosis, ovarian cancer, and uterine fibroids. The high NPD tertile (3) resulted in a significantly earlier diagnosis than the other tertiles for breast cancer, endometriosis, ovarian cancer, and uterine fibroids. No phenotypes had three out of three significant comparisons between the SL tertiles, but the highest SL tertile was associated with earlier diagnosis of endometriosis, while the lowest SL tertile was associated with a later diagnosis of uterine fibroids.

Next, we performed the same time-to-event analyses for the lifestyle variables: AU, SK, ST, and SM (Figure 4). The different AU tertile groups didn’t show significant differences for age at diagnosis, except for between the first and second tertiles in breast cancer (P = 2.2 x 10^{-3}); those who drink lightly get diagnosed with breast cancer earlier than those that drink moderately. Similarly, different levels of sedentary minutes also didn't significantly impact diagnosis except for between the first and third tertiles in breast cancer (P = 4.4 x 10^{-2}), with those in the high SM curve, get diagnosed later than the low SM group. Smokers in the third tertile get diagnosed with uterine fibroids earliest (P vs Low = 2.3 x 10^{-3}, P vs Medium = 1.8 x 10^{-11}). Breast cancer cases in the lowest tertile of steps get diagnosed latest (P vs Medium = 8.6x10^{-5}, P vs High = 1.4x10^{-2}), this could be confounded by age as older women likely take fewer daily steps. For preeclampsia and uterine cancer cases, those in the third tertile of steps get diagnosed latest.

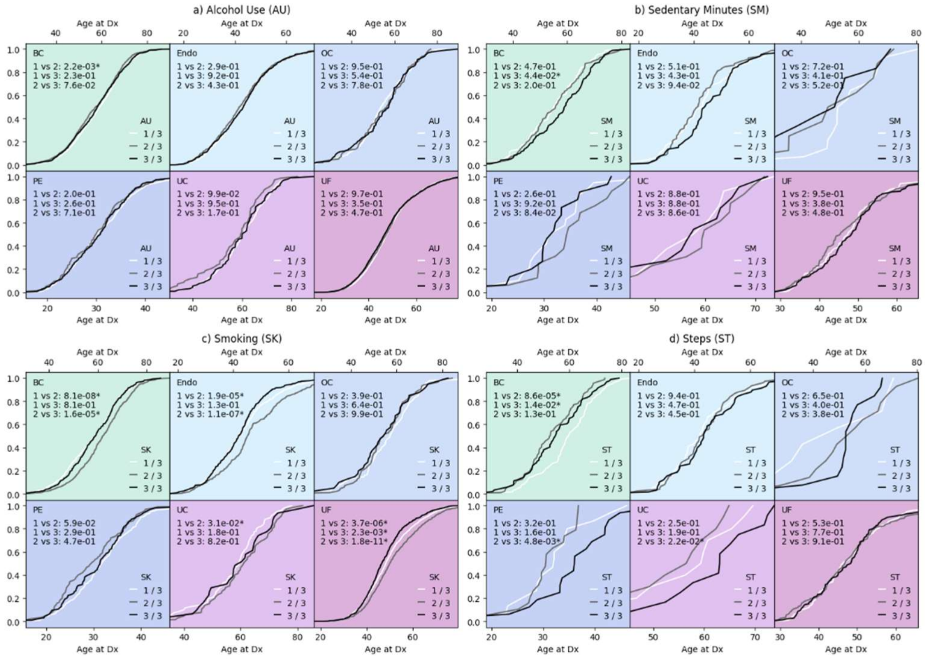
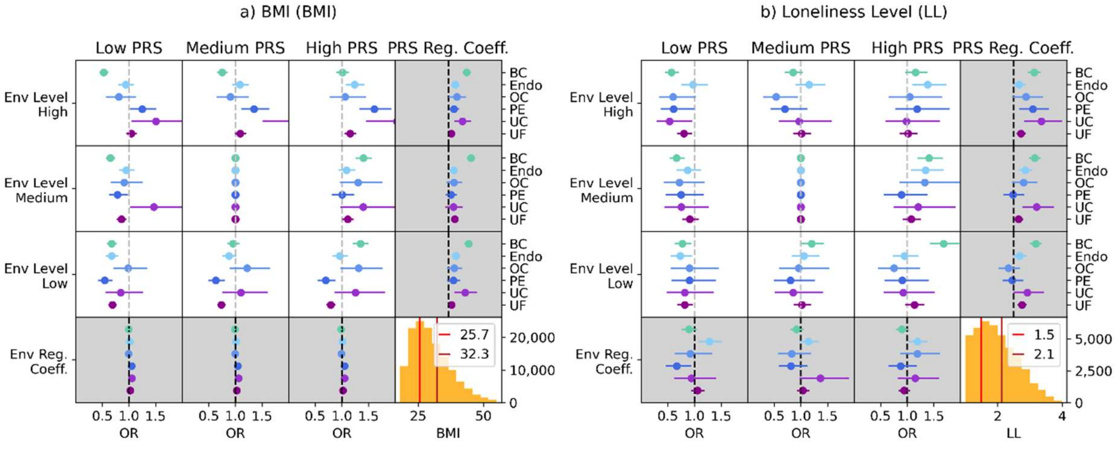


Figure 4: time-to-event analyses for lifestyle measurements (a - alcohol use, b - sedentary minutes, c - smoking, and d - steps). Each panel shows three "survival" curves per phenotype, stratified by the value of the environmental measure where 1 is the lowest tertile and 3 is the highest tertile. The x-axes represent age at diagnosis (Dx). Also indicated in each grid cell are the p-values of pairwise log rank comparisons between those three curves. Any p-values less than 0.05 are annotated with an asterisk. BC: Breast Cancer; UF: Uterine Fibroids; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia

3.4 *Genetic risk effects vary by environmental context*

We assigned every individual to a genetic risk tertile (low, medium, high) and an environmental exposure level (low, medium, high), the combinations of which resulted in nine sub-groups. Within each of the sub-groups, we computed the odds ratio of the phenotype relative to the medium-medium group. We also performed stratified logistic regressions to estimate the PRS and environmental measurement effects. Because NPD and NSD scores were highly correlated, we only tested NPD. First, we focused on the three remaining SDoH and BMI (Figure 5).



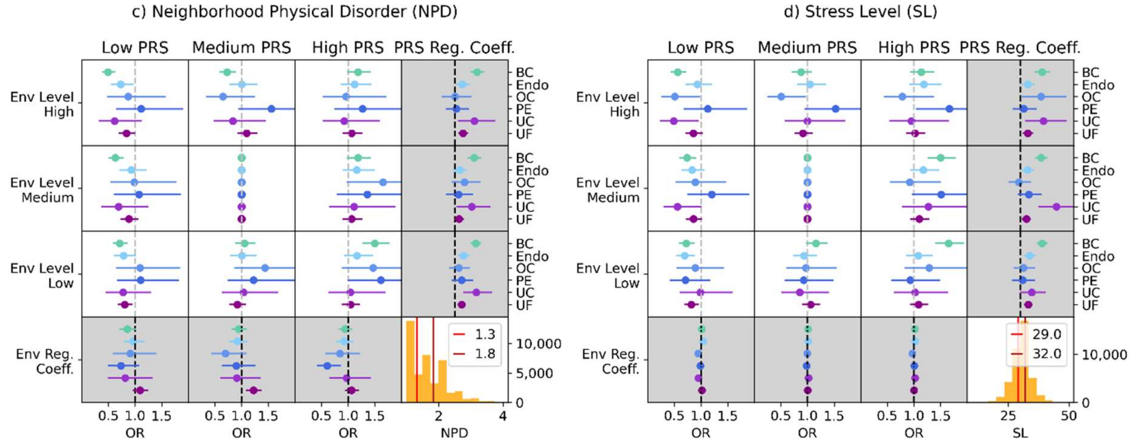
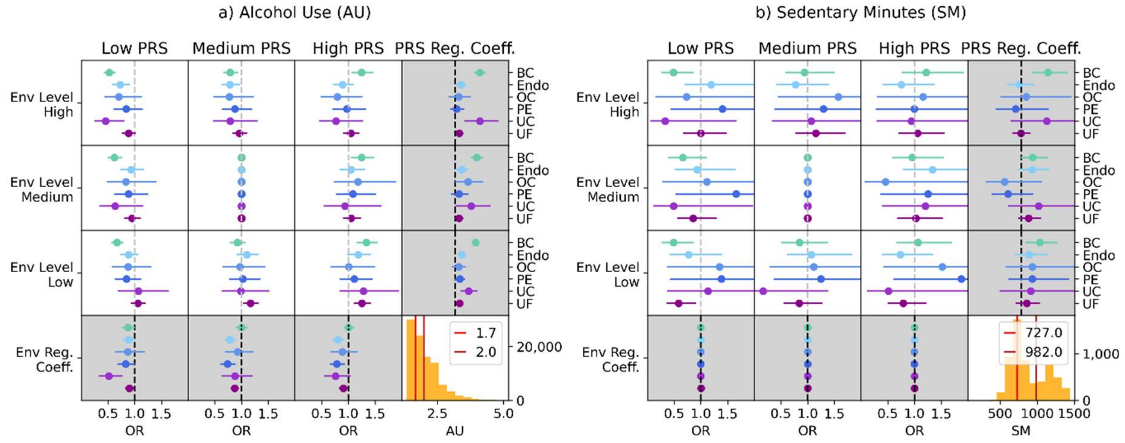


Figure 5: All odds ratio and logistic regression tests performed for BMI and SDoH. The environmental factors are (a) BMI, (b) loneliness, (c) neighborhood physical disorder, and (d) stress. The upper left 3x3 grid in each pane shows the odds ratios of the phenotypes in each cell. The rightmost column shows regression coefficients stratified by environmental tertile. The bottom row shows regression coefficients stratified by genetic risk. The bottom right cell shows a histogram of the environmental variable, with the cutoffs between the tertiles marked. BC: Breast Cancer; UF: Uterine Fibroids; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia

The BMI tertiles were split at 25.7 and 32.3, which are near the conventional cutoffs for overweight (25) and obese (30). At all levels of genetic risk (low, medium, and high), BMI was positively associated with preeclampsia, uterine cancer, and uterine fibroids. BMI was negatively associated with breast cancer. Chronic loneliness and stress are known to be detrimental to long-term health. In the lowest genetic risk group, loneliness was positively associated with endometriosis. Those in the medium and high loneliness groups were more susceptible to genetic risk of ovarian cancer, preeclampsia, and uterine cancer.

Next, we focused on modulating effects of lifestyle factors, including the two Fitbit variables, smoking, and alcohol use (Figure 6).



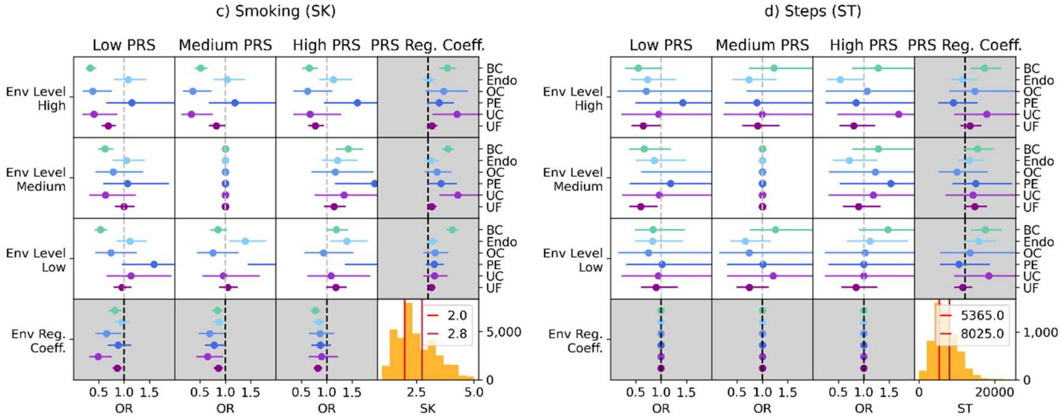


Figure 6: All odds ratio and logistic regression tests performed for the lifestyle variables. The environmental factors are (a) alcohol use, (b) sedentary minutes, (c) smoking, and (d) steps. The upper left 3x3 grid in each pane shows the odds ratios of the phenotypes in each cell. The rightmost column shows regression coefficients stratified by environmental tertile. The bottom row shows regression coefficients stratified by genetic risk. The bottom right cell shows a histogram of the environmental variable, with the cutoffs between the tertiles marked. BC: Breast Cancer; UF: Uterine Fibroids; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia

AU had a highly skewed distribution, so the cutoffs between the three tertiles were close together (1.7 vs 2.0). The effect sizes of the PRSs for breast cancer, endometriosis, and uterine cancer were strongest in the tertile with the highest drinking scores. Notably, SK had an inverse effect on breast cancer and uterine fibroids at all levels of genetic risk. Since the models were adjusted for age, it is unlikely that age is confounding these results. Additionally, within the lowest smoking group, the PRS coefficient was not significant, but it was significant for the medium and high smokers. SM had a bimodal distribution. Due to the smaller sample size of the Fitbit data, most of the test statistics were not significant. However, the breast cancer PRS was significantly associated with breast cancer for those who were the most sedentary. Similarly, most of the effect sizes for the steps tests were not significant, but the effect of the breast cancer PRS was significant in the group that took the fewest daily steps on average.

4 Discussion

In this study, we evaluated the effects of environmental variables on women's health outcomes. Specifically, we looked at effects on age at diagnosis and modulation of genetic risk. In 145,563 women in AOU, we analyzed six risk models for women's health diseases. From there, we calculated stratified effect sizes for each PRS for tertiles of each environmental measurement. Overall, we showed that genetic risk models are significantly impacted by different environmental contexts. In general, the most severely affected group of the environment had the strongest effect of the PRS and often resulted in the earliest. These findings underscore the necessity of integrating diverse environmental and social factors into disease risk models to capture the full spectrum of influences on health.

Of the 12 PRSs tested based on their performance in the PGS catalog, nine showed significant positive associations with their respective phenotypes, with breast cancer demonstrating the strongest association. The disparity between the sample population used to create these risk scores and the AOU biobank likely influenced these results, as PRS performance is highly sensitive to population mismatch[41]. There were differences between the derivation datasets and AOU's unique composition, with about half of the genomic dataset comprising participants of non-European ancestry[42]. This highlights a key drawback of existing PRSs, which are often based on European populations, limiting their relevance for non-European individuals. Notably, genetic risk did not significantly affect age at diagnosis for the six best risk scores, aligning with expectations, as these scores were derived from studies evaluating lifetime disease risk rather than onset. Factors such as SDoH and environmental influences, often correlated with race and ancestry, also play a role in disease susceptibility.

BMI has been significantly associated with a multitude of gynecological conditions[43]. In the current study, we have demonstrated that high BMI can serve as a risk factor for earlier diagnosis of breast, ovarian, and uterine cancer as well as uterine fibroids. Furthermore, BMI was found to be associated with preeclampsia, uterine cancer and uterine fibroids, across all genetic risk groups. Preeclampsia is a pregnancy-related condition, so it is possible that several of the environmental risk factor measurements (BMI, activity levels) may not be representative of the woman's environment at the time of onset as these variables are affected by pregnancy. However, we aimed to evaluate average lifestyle trends, including time leading up to pregnancy. These findings, in conjunction with previous reports on metabolism-related genes on various female cancer types[44,45], emphasize the importance of incorporating environmental factors, especially BMI, for a holistic understanding of disease risk and health outcomes.

The lowest genetic risk groups for endometriosis, preeclampsia, ovarian cancer, and uterine cancer showed positive associations at multiple levels of loneliness. This highlights the profound impact that social and psychological factors can have on physical health. By considering and stratifying risk factors based on both genetic and environmental factors, we can potentially facilitate earlier detection of health burden across diverse population groups. It allows us to identify individuals who, despite having a low genetic risk, may still be at high overall risk due to adverse environmental or social conditions, and ultimately enhance health outcomes for a broader spectrum of the population.

Our study has several limitations. One limitation is that EHR-based phenotyping can be challenging for complex disorders, especially in women's health diseases which are often under-diagnosed, such as uterine fibroids[46] and endometriosis[47]. Phenotyping algorithms have been previously designed to compute phenotypes more accurately than ICD codes alone. Their use in our study is restricted by reliance on clinical notes[48], which are not available in AOU. Other large genomic biobank studies, have leveraged ICD- or PheCode-based case-control phenotyping[1,49,50]. While the accuracy of ICD codes alone varies across the phenotypes, a key advantage of large biobank data is that the substantial sample size can help mitigate the impact of noise introduced by imprecise phenotyping, leading to more robust statistical associations[51].

Another limitation of our study was that we used age at the first diagnosis code of a condition as a proxy for disease onset. Depending on how patients move between healthcare systems, a common occurrence in the EHR is that a condition may have been diagnosed earlier at a different facility, but the corresponding diagnosis code is entered into the EHR only after the patient joins a new healthcare system. This introduces potential noise into the age variable, as the true onset might have been recorded elsewhere or at a different time. However, since many of our sample sizes were large enough to yield significant effects, which should have counteracted the noise. We found that higher-risk environmental groups typically had earlier diagnoses. Given the EHR data, it can be hard to disentangle earlier diagnosis due to earlier onset versus earlier diagnosis due to increased vigilance based on existing risk factors.

Survey data are notoriously challenging to work with, so we were also limited by potential noise introduced by the self-reporting process. To mitigate error, we divided the participants into subgroups by environmental variable tertiles rather than relying on the exact quantitative measures. However, stratifying the individuals into subgroups reduced the sample size and statistical power for each regression. The observations that smoking levels seemed to have non-monotonic effects (medium smokers get diagnosed later with breast cancer, endometriosis, and uterine fibroids) may stem from confounders in the survey measurements. Our overall approach, though it has a few limitations, has provided a practical and scalable way to examine multi-modal predictive and progression models of women's health diseases.

Due to systemic challenges faced by marginalized communities, such populations find themselves exposed to environmental stressors at greater rates[52]. Differing odds ratios for those with similar levels of genetic risk but different levels of environmental risk suggest that not including environmental risk factors in predictive models utilizing PRS could lead to inaccurate risk assessments and potentially overlook significant contributors to disease susceptibility. The current study identifies the dangers in reductionist approach to disease stratification and risk prediction, based solely on either genetics or environmental factors. This suggests that integrating both the genetic and environmental components into a specific disease model would help better classify individual risk.

In the future, using nonlinear approaches for risk modeling which capture variable interactions such as multilayer perceptron could aid in more accurately representing complex relationships between genetics, environmental risk factors, and the phenotypes. While those types of models are harder to train, we can now take advantage of growing data repositories, including AOU, to develop generalizable models that capture important modalities of risk variables. We included eight environmental risk factors, four SDoH and four lifestyle measurements, which capture some, but not all, external influences. Future methodologies may include more risk factors but also should account for potential missing data, as it can be challenging to administer surveys and/or collect wearables data on a large scale. In the future, we also hope to replicate these results in additional biobanks.

Complex systems approaches to incorporate multi-directional interactions between patients and their environment, such as those modeled here, are better suited to leverage the power of genomic data in making widely applicable, clinically relevant tools. Further attempts to strengthen the predictive ability of PRS models need not focus solely on improving the identification of relevant loci, but also relevant environmental risk factors including SDoH. By improving our understanding and application of PRSs, especially in underrepresented areas like women's health, we can enhance disease prediction, prevention, and personalized treatment strategies.

5 Acknowledgments

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.

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number R01HD110567.

6 References

1. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022).
2. Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024).
3. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
4. Mirin, A. A. Gender Disparity in the Funding of Diseases by the U.S. National Institutes of Health. *Journal of Women's Health* **30**, 956–963 (2021).
5. Schubert, K. G., Bird, C. E., Kozhimmanil, K. & Wood, S. F. To Address Women's Health Inequity, It Must First Be Measured. *Health Equity* **6**, 881–886 (2022).
6. Shah, P. D. Polygenic Risk Scores for Breast Cancer—Can They Deliver on the Promise of Precision Medicine? *JAMA Network Open* **4**, e2119333 (2021).
7. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
8. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
9. Pulley, J., Clayton, E., Bernard, G. R., Roden, D. M. & Masys, D. R. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin Transl Sci* **3**, 42–48 (2010).
10. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology* **27**, S2–S8 (2017).
11. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *Journal of Personalized Medicine* **12**, 1974 (2022).
12. The "All of Us" Research Program. *New England Journal of Medicine* **381**, 668–676 (2019).

13. Eklund, M. *et al.* The WISDOM Personalized Breast Cancer Screening Trial: Simulation Study to Assess Potential Bias and Analytic Approaches. *JNCI Cancer Spectr* **2**, pky067 (2019).

14. Lennon, N. J. *et al.* Selection, optimization, and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse populations. *medRxiv* 2023.05.25.23290535 (2023) doi:10.1101/2023.05.25.23290535.

15. Zhang, Y. & Ma, N.-Y. Environmental Risk Factors for Endometriosis: An Umbrella Review of a Meta-Analysis of 354 Observational Studies With Over 5 Million Populations. *Front. Med.* **8**, (2021).

16. Daly, A. A., Rolph, R., Cutress, R. I. & Copson, E. R. A Review of Modifiable Risk Factors in Young Women for the Prevention of Breast Cancer. *Breast Cancer: Targets and Therapy* **13**, 241–257 (2021).

17. Vafaei, S., Alkhrait, S., Yang, Q., Ali, M. & Al-Hendy, A. Empowering Strategies for Lifestyle Interventions, Diet Modifications, and Environmental Practices for Uterine Fibroid Prevention; Unveiling the LIFE UP Awareness. *Nutrients* **16**, 807 (2024).

18. Sundström, A., Adolfsson, A. N., Nordin, M. & Adolfsson, R. Loneliness Increases the Risk of All-Cause Dementia and Alzheimer's Disease. *The Journals of Gerontology: Series B* **75**, 919–926 (2020).

19. Ajibewa, T. A. *et al.* Chronic Stress and Cardiovascular Events: Findings From the CARDIA Study. *American Journal of Preventive Medicine* **67**, 24–31 (2024).

20. Crear-Perry, J. *et al.* Social and Structural Determinants of Health Inequities in Maternal Health. *Journal of Women's Health* **30**, 230–235 (2021).

21. Katon, J. G., Plowden, T. C. & Marsh, E. E. Racial disparities in uterine fibroids and endometriosis: a systematic review and application of social, structural, and political context. *Fertility and Sterility* **119**, 355–363 (2023).

22. Kurani, S. S. *et al.* Association of Neighborhood Measures of Social Determinants of Health With Breast, Cervical, and Colorectal Cancer Screening Rates in the US Midwest. *JAMA Network Open* **3**, e200618 (2020).

23. Kim, S. *et al.* A comprehensive gene–environment interaction analysis in Ovarian Cancer using genome-wide significant common variants. *International Journal of Cancer* **144**, 2192–2205 (2019).

24. Domingue, B. W., Trejo, S., Armstrong-Carter, E. & Tucker-Drob, E. M. Interactions between Polygenic Scores and Environments: Methodological and Conceptual Challenges. *Sociol Sci* **7**, 465–486 (2020).

25. Musliner, K. L. *et al.* Polygenic liability, stressful life events and risk for secondary-treated depression in early life: a nationwide register-based case-cohort study. *Psychological Medicine* **53**, 217–226 (2023).

26. Data Browser | All of Us Public Data Browser. https://databrowser.researchallofus.org/.

27. Hallinan, C. M. *et al.* Seamless EMR data access: Integrated governance, digital health and the OMOP-CDM. *BMJ Health Care Inform* **31**, e100953 (2024).

28. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* **53**, 420–425 (2021).

29. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).

30. Tesfaye, S. *et al.* Measuring social determinants of health in the All of Us Research Program. *Sci Rep* **14**, 8815 (2024).

31. Park, J. H., Moon, J. H., Kim, H. J., Kong, M. H. & Oh, Y. H. Sedentary Lifestyle: Overview of Updated Evidence of Potential Health Risks. *Korean J Fam Med* **41**, 365–373 (2020).

32. Inoue, K., Tsugawa, Y., Mayeda, E. R. & Ritz, B. Association of Daily Step Patterns With Mortality in US Adults. *JAMA Network Open* **6**, e235174 (2023).

33. Rich, J. T. *et al.* A practical guide to understanding Kaplan-Meier curves. *Otolaryngol Head Neck Surg* **143**, 331–336 (2010).

34. Davidson-Pilon, C. lifelines: survival analysis in Python. *Journal of Open Source Software* **4**, 1317 (2019).

35. Shieh, Y. *et al.* Development and testing of a polygenic risk score for breast cancer aggressiveness. *npj Precis. Onc.* **7**, 1–11 (2023).

36. Tanigawa, Y. *et al.* Significant sparse polygenic risk scores across 813 traits in UK Biobank. *PLOS Genetics* **18**, e1010105 (2022).

37. Dareng, E. O. *et al.* Polygenic risk modeling for prediction of epithelial ovarian cancer risk. *Eur J Hum Genet* **30**, 349–362 (2022).

38. Piekos, J. A. *et al.* Uterine fibroid polygenic risk score (PRS) associates and predicts risk for uterine fibroid. *Hum Genet* **141**, 1739–1748 (2022).

39. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics* **109**, 12–23 (2022).

40. Kloeve-Mogensen, K. *et al.* Polygenic Risk Score Prediction for Endometriosis. *Frontiers in Reproductive Health* **3**, (2021).

41. Wang, Y. *et al.* Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. *Cell Genomics* **3**, (2023).

42. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).

43. Venkatesh, S. S. *et al.* Obesity and risk of female reproductive conditions: A Mendelian randomisation study. *PLOS Medicine* **19**, e1003679 (2022).

44. Hua, Y., Gao, L. & Li, X. Comprehensive Analysis of Metabolic Genes in Breast Cancer Based on Multi-Omics Data. *Pathol Oncol Res* **27**, 1609789 (2021).

45. M, M., Tj, R.-F., A, K. & Rj, S. Genetics of enzymatic dysfunctions in metabolic disorders and cancer. *Frontiers in oncology* **13**, (2023).

46. Ahmad, A. *et al.* Diagnosis and management of uterine fibroids: current trends and future strategies. *Journal of Basic and Clinical Physiology and Pharmacology* **34**, 291–310 (2023).

47. Soliman, A. M., Fuldeore, M. & Snabes, M. C. Factors Associated with Time to Endometriosis Diagnosis in the United States. *Journal of Women's Health* **26**, 788–797 (2017).

48. Hoffman, S. R. *et al.* Optimizing research in symptomatic uterine fibroids with development of a computable phenotype for use with electronic health records. *American Journal of Obstetrics and Gynecology* **218**, 610.e1-610.e7 (2018).

49. GWAS and enrichment analyses of non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE Network | SpringerLink. https://link.springer.com/article/10.1186/s12916-019-1364-z.

50. Rahmioglu, N. *et al.* The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. *Nat Genet* **55**, 423–436 (2023).

51. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102–1111 (2013).

52. Evans, G. W. & Kantrowitz, E. Socioeconomic Status and Health: The Potential Role of Environmental Risk Exposure. *Annual Review of Public Health* **23**, 303–331 (2002).