International Journal of
Environmental Research and Public Health

MDPI

Article

Estimating the Relative Contribution of Environmental and Genetic Risk Factors to Different Aging Traits by Combining Correlated Variables into Weighted Risk Scores

Claudia Wigmann [1], Anke Hüls [2,3], Jean Krutmann [1,4] and Tamara Schikowski [1,*]

[1] IUF—Leibniz Research Institute for Environmental Medicine, 40225 Duesseldorf, Germany
[2] Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA
[3] Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA
[4] The Human Phenome Institute, Shanghai 200433, China
* Correspondence: tamara.schikowski@iuf-duesseldorf.de; Tel.: +49-211-3389-341

Abstract: Genetic and exposomal factors contribute to the development of human aging. For example, genetic polymorphisms and exposure to environmental factors (air pollution, tobacco smoke, etc.) influence lung and skin aging traits. For prevention purposes it is highly desirable to know the extent to which each category of the exposome and genetic factors contribute to their development. Estimating such extents, however, is methodologically challenging, mainly because the predictors are often highly correlated. Tackling this challenge, this article proposes to use weighted risk scores to assess combined effects of categories of such predictors, and a measure of relative importance to quantify their relative contribution. The risk score weights are determined via regularized regression and the relative contributions are estimated by the proportion of explained variance in linear regression. This approach is applied to data from a cohort of elderly Caucasian women investigated in 2007–2010 by estimating the relative contribution of genetic and exposomal factors to skin and lung aging. Overall, the models explain 17% (95% CI: [9%, 28%]) of the outcome's variance for skin aging and 23% ([11%, 34%]) for lung function parameters. For both aging traits, genetic factors make up the largest contribution. The proposed approach enables us to quantify and rank contributions of categories of exposomal and genetic factors to human aging traits and facilitates risk assessment related to common human diseases in general. Obtained rankings can aid political decision making, for example, by prioritizing protective measures such as limit values for certain exposures.

Keywords: aging; environmental exposure; exposome; relative contribution; relative importance; risk score

Citation: Wigmann, C.; Hüls, A.; Krutmann, J.; Schikowski, T. Estimating the Relative Contribution of Environmental and Genetic Risk Factors to Different Aging Traits by Combining Correlated Variables into Weighted Risk Scores. *Int. J. Environ. Res. Public Health* **2022**, *19*, 16746. https://doi.org/10.3390/ijerph192416746

Academic Editor: Jimmy T. Efird

Received: 9 November 2022
Accepted: 10 December 2022
Published: 13 December 2022

1. Introduction

Human phenotypes in general and health outcomes such as aging traits in particular result from genetic and non-genetic influences. For the latter the term exposome has been coined, which according to Christopher Wild is the totality of all non-genetic factors a human individual is exposed to from conception to death [1]. The exposome not only encompasses environmental factors but also lifestyle and behaviors, social environment and social status as well as the biological response [2].

The exposome concept is a step towards an all-embracing assessment of environment and health. It is potentially very interesting with regards to risk assessment, because knowledge about the relative contribution of distinct exposomal and genetic factors to a specific health outcome or aging trait would allow for more a precise and efficient prevention.

However, to date it is still challenging to measure all exposures of the exposome continuously over the entire lifetime for each individual [3]. Apart from the difficulty in measuring the exposome, the statistical analysis is also not straightforward [3]. Even a

cross-sectional analysis of the complex associations between different exposures, genetics and health outcomes is a statistical challenge, as many exposures are highly correlated. Epidemiological studies typically report exposure–health associations on an individual exposure-by-exposure basis while adjusting for confounding [3]. Correlated exposures are usually not included in traditional regression models due to concerns about multicollinearity. A recent review of statistical approaches used in the context of exposome studies is given in Guillien et al. (2021) [4]. Most of these approaches focus on the detection of causal exposures, i.e., variable selection, and on prediction of individual health status.

This article provides means for studying the contribution of a risk factor category such as air pollution as a whole, that is, the combined contribution of different exposures (e.g., nitrogen dioxide and particulate matter) belonging to the same category. The goal of the here-proposed methodology is to determine the extent to which each of several categories within the exposome as well as genetic factors contribute to a certain health outcome. Knowing these relative contributions would also enable comparisons between different populations and between different outcomes.

Since the concept of the exposome was introduced as a complement to the genome, methodology developed for genome-wide analyses might also be useful for analyses of the exposome. Consequently, exposome-wide association studies (in analogy to genome-wide association studies) have been proposed to assess the exposome [5]. These association studies use univariate single-exposure regression that does not take co-exposures into account and hence cannot provide a measure of relative contribution of a whole category of risk factors.

Approaches for multi-exposure regression like penalized regression models (LASSO [6] or elastic net [7]), the deletion/substitution/addition (DSA) algorithm [8], weighted quantile sum regression [9] (and its extension "quantile-based g-computation" [10]) or Bayesian kernel machine regression [11] take co-exposures into account, but are not useful for estimating the relative contributions of several risk factor categories either. Penalized regression and the DSA algorithm are variable selection methods and cannot directly infer the combined effects of categorized exposures. Weighted quantile sum regression and quantile-based g-computation combine several correlated exposures into one index and enable the estimation of the joint effect of an exposure mixture, but cannot handle categorical exposures. In addition, an extension for several risk factor categories of the exposome with more than one index remains unclear. Bayesian kernel machine regression can take hierarchical structures of the exposome into account by partitioning correlated exposures into several categories ("groups") through a-priori information, but only a single exposure per category and not a mixture is allowed to contribute to the final estimates. In addition, this method is also not able to handle categorical exposures.

A multivariate approach for genome-wide analyses is polygenic risk scores [12], which was developed for determining the genetic basis underlying a trait or disease, where the genetic predictors are in part highly correlated. Here, we will borrow the polygenic risk score methodology and adopt it to the exposome concept. Weighted risk scores, defined as weighted sums of all exposures belonging to the same risk factor category, will be constructed to analyze the combined impact of different (correlated) environmental exposures on health. More precisely, the proposed methodology consists of two steps: (1) combining correlated predictors into one weighted risk score (RS) per risk factor category and (2) estimating the relative contribution of each RS to a certain outcome using shares of explained variance in a linear regression model. The weights of the predictors will be determined internally in a training sample using regularized regression (explicitly, ridge regression) and a bootstrapping approach to account for the randomness in splitting the data set into training and test set. Estimating the weights by regularized regression in a training sample has been widely used for the construction of polygenic risk scores in general [13,14] and in the context of gene–environment–interaction studies [15–17]. In the second step, the Lindeman–Merenda–Gold measure of relative importance [18] is applied to assess the contribution of the composed risk scores to the health outcome in the test sample.

Here, data from the German SALIA cohort (Study on air pollution, lung function, inflammation and aging) is analyzed to prove the concept of applying risk scores to estimate relative contributions of environmental and genetic risk factors. The concept is applied to two aging traits, namely skin aging and aging-associated decreased lung function, to evaluate the relative contributions. The focus is on these two aging traits because for both it is well established that genetic and a number of specific exposomal factors contribute to their development [19,20]. Accordingly, important exposomal factors for lung aging are tobacco smoke and air pollution [20–22]; for skin aging these include exposure to ultraviolet (UV) radiation, air pollution and tobacco smoke, as well as nutritional factors [19].

2. Materials and Methods

The proposed methodology was applied to data from the SALIA cohort study of elderly German women by analyzing three aging traits: a z-score of facial pigment spots and z-scores of the lung function parameters Forced Expiratory Volume in 1 s (FEV_1) and Forced Vital Capacity (FVC). Details on the study population [23–25] as well as outcome [26–29], genetic [30–36], exposure [37–40] and confounder variables can be found in the Supplementary Methods and Supplementary Tables S1–S3 (Supplementary File S1). An overview of the variables is given in Figure 1 and descriptive statistics can be found in Table 1. The de-correlating effect of building the risk scores is demonstrated with correlation plots in Supplementary Figures S1–S5 (Supplementary File S1).

Table 1. Descriptive statistics of all outcome and predictor variables in both analysis samples.

Variable	Categories	Mean (SD) or *n* (%)	
		Skin aging analysis (n = 547)	Lung function analysis (n = 510)
Outcomes			
Pigment spot z-score		0.00 (1.00)	
FEV_1 z-score			0.17 (1.02)
FVC z-score			0.24 (0.90)
FEV_1/FVC z-score			−0.22 (0.85)
Single Predictors			
Age [years]		73.58 (2.97)	73.52 (2.99)
Height [cm]			162.83 (5.76)
SES:	1 = "<10 years"	94 (17.2%)	89 (17.5%)
	2 = "10 years"	275 (50.3%)	249 (48.8%)
	3 = ">10 years"	178 (32.5%)	172 (33.7%)
Skin type:	dark (0; Fitzpatrick type 3 and 4)	243 (44.4%)	
	light (1; Fitzpatrick type 1 and 2)	304 (55.6%)	
Sunbed use:	never (0)	459 (83.9%)	
	ever (1)	88 (16.1%)	
SPF:	no (0)	214 (39.1%)	
	yes (1)	333 (60.9%)	
HRT:	no (0)	319 (58.3%)	
	yes (1)	228 (41.7%)	
MeDi score		28.50 (2.79)	
Obesity Risk Score			
mean BMI up to FU 1 [kg/m^2]		26.70 (3.67)	26.66 (3.66)
BMI FU 2 [kg/m^2]		27.32 (4.34)	27.29 (4.34)
Lack of physical activity:	no (0)	216 (39.5%)	203 (39.8%)
	yes (1)	331 (60.5%)	307 (60.2%)
Smoking Risk Score			
Current smoking:	no (0)	535 (97.8%)	498 (97.6%)
	yes (1)	12 (2.2%)	12 (2.4%)
Former smoking:	no (0)	458 (83.7%)	425 (83.3%)
	yes (1)	89 (16.3%)	85 (16.7%)

Table 1. *Cont.*

Variable	Categories	Mean (SD) or *n* (%)	
		Skin aging analysis (n = 547)	Lung function analysis (n = 510)
ETS at work:	never (0)	314 (57.4%)	294 (57.6%)
	ever (1)	233 (42.6%)	216 (42.4%)
ETS at home:	never (0)	365 (66.7%)	337 (66.1%)
	ever (1)	182 (33.3%)	173 (33.9%)
Packyears [packs/day × years]		3.68 (12.46)	3.83 (12.74)
Air pollution Risk Score			
Residence:	rural (0)	260 (47.5%)	249 (48.8%)
	urban (1)	287 (52.5%)	261 (51.2%)
NO_2 [$\mu g/m^3$]		37.73 (11.46)	37.21 (11.16)
NO_x [$\mu g/m^3$]		70.52 (32.66)	69.17 (31.93)
PM_{10} [$\mu g/m^3$]		49.27 (7.17)	48.98 (7.38)
$PM_{2.5}$ [$\mu g/m^3$]		32.75 (4.65)	32.56 (4.78)
PM_{coarse} [$\mu g/m^3$]		17.57 (3.85)	17.42 (3.91)
$PM_{2.5abs}$ [$10^{-5}/m$]		2.74 (0.92)	2.71 (0.92)
Trafloadmajor [1000 vehicles × m/day]		900.63 (2313.15)	839.51 (2200.63)
Invdistmajor [1/m]		0.01 (0.02)	0.01 (0.02)
Radiation Risk Score			
UV-B [J/m^2]		3140.38 (33.03)	
UV index [40 W/m^2]		7.18 (0.09)	

BMI: body mass index; ETS: environmental tobacco smoke; FEV_1: forced expiratory volume in 1 s; FVC: forced vital capacity; FU: follow-up; HRT: hormone replacement therapy; Invdistmajor: inverse distance to next major road (>5000 vehicles/day); MeDi: Mediterranean diet (definition: see Supplementary Materials); n: number of samples; NO_2: nitrogen dioxide; NO_x: nitrogen oxides; PM_{10}: particulate matter with aerodynamic diameter $\leq$ 10 µm, $PM_{2.5}$: particulate matter with aerodynamic diameter $\leq$2.5 µm; PM_{coarse}: coarse fraction of PM_{10} calculated as PM_{10} minus $PM_{2.5}$; $PM_{2.5abs}$: absorbance of particulate matter with aerodynamic diameter of $\leq$2.5 µm; SD: standard deviation; SES: socio-economic status; SPF: sun protection factor; Trafloadmajor: total traffic load (number of vehicles/day × length of road segments) from major roads (>5000 vehicles/day) within 100 m buffer; UV: ultraviolet.

2.1. Choice of Predictors

The risk scores combinepredefined predictors. In addition to the risk scores, further predictors are included as single predictors in the linear regression models. Figure 1 gives an overview for both the skin aging and the lung function analyses.

All predictors except the SNP variables have been standardized to mean zero and standard deviation one to achieve a fair penalization of all regressors in the following.

2.2. Bootstrap Data Sets and Division into Training and Test Sample

The following analysis steps are repeated B times using the bootstrap principle. That means that the analysis is not only conducted on the original data set, but also on *B-1* data sets created from the original data set by randomly sampling participants with replacement. Each bootstrap data set is then divided randomly into training and test samples in the relation 60% to 40%, as recommended in [12]. The number of bootstrap data sets is set to B = 500 for the skin aging outcome and B = 200 for the lung function outcomes, since in the latter case 278 SNP variables are included in the genetic RS resulting in considerably longer computation times.

2.3. Learning Risk Score Weights on Training Sample

The weights for the risk scores were learned on 60% of the participants in the training sample using ridge regression (implemented by elastic net regression with parameter $\alpha = 0$), where the regularization parameter lambda was chosen via tenfold cross-validation with the R function cv.glmnet from R package glmnet [41]. Ridge regression was chosen, since variable selection was not the ultimate goal and shrinking all coefficients towards zero

should retain the relations between the variables and lead to meaningful RS weights. The model formula, exemplary for the lung function analyses, is given by

$$y = \gamma_0 + \underbrace{\sum_{j=1}^{4} \gamma_{1,j}\, x_{1,j}}_{\text{Single predictors}} + \underbrace{\sum_{j=1}^{278} \gamma_{2,j}\, x_{2,j}}_{\text{Genetic predictors}} + \underbrace{\sum_{j=1}^{3} \gamma_{3,j}\, x_{3,j}}_{\text{Obesity predictors}} + \underbrace{\sum_{j=1}^{5} \gamma_{4,j}\, x_{4,j}}_{\text{Smoking predictors}} + \underbrace{\sum_{j=1}^{9} \gamma_{5,j}\, x_{5,j}}_{\text{Air pollution predictors}}.$$

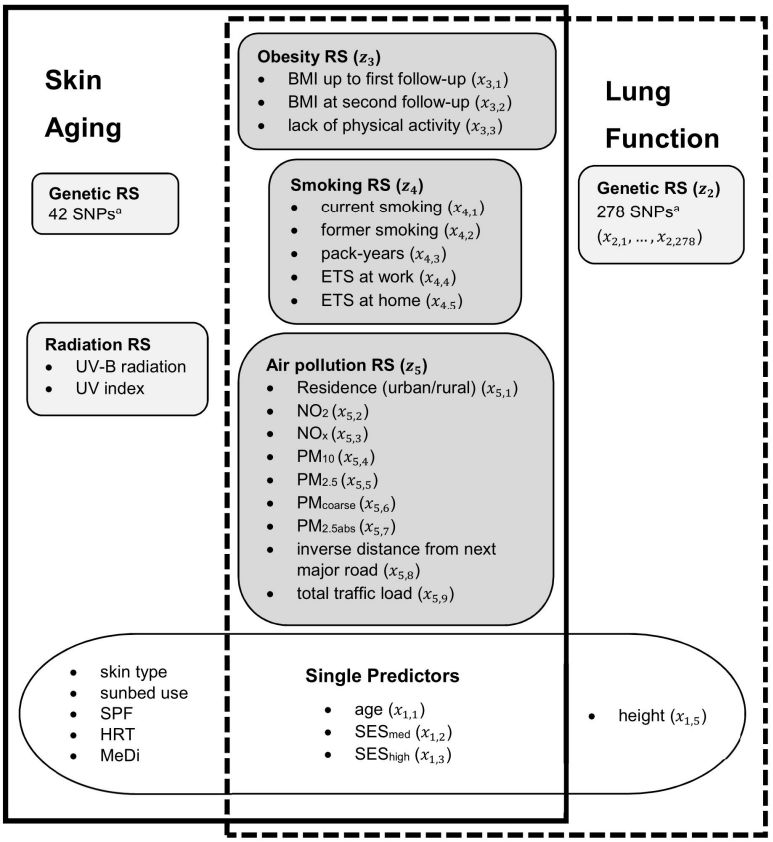

Figure 1. Overview of the risk scores and the combined predictors for both aging outcomes. The notation used in the description of the statistical approach is given in brackets, exemplary for lung function. BMI: body mass index; ETS: environmental tobacco smoke; HRT: hormone replacement therapy; MeDi: Mediterranean diet (definition: see Supplementary Materials); NO_2: nitrogen dioxide; NO_x: nitrogen oxides; PM_{10}: particulate matter with aerodynamic diameter $\leq$ 10 µm; $PM_{2.5}$: particulate matter with aerodynamic diameter $\leq$ 2.5 µm; PM_{coarse}: coarse fraction of PM_{10} calculated as PM_{10} minus $PM_{2.5}$; $PM_{2.5abs}$: absorbance of particulate matter with aerodynamic diameter of $\leq$2.5 µm; RS: risk score; SES: socio-economic status; SNP: single nucleotide polymorphism; SPF: sun protection factor; UV: ultraviolet. [a] Details on the selected SNPs are given in Supplementary Tables S2 and S3 (Supplementary File S1).

The single predictors are included in this regression model to account for their effects on the outcome, but only the predictors belonging to the risk scores are regularized, since they are highly correlated and their coefficients will be used as weights in the risk scores. Since the folds for the cross-validation in cv.glmnet are selected at random, the results are random as well. To reduce this randomness, the estimation of the coefficients $\gamma_{k,j}$ is repeated twenty times and the resulting estimates are averaged. For each RS the respective subset of coefficients are normalized so that the resulting weights lie in $[-1, 1]$ and sum

to one for each RS. Explicitly, the weights of the risk scores are (exemplary for the lung function analyses) calculated as

$$w_{k,j} = \begin{cases} \frac{\hat{\gamma_{k,j}}}{\sum_{l=1}^{m_k} |\hat{\gamma_{l,j}}|}, & \exists\, j\ \hat{\gamma_{k,j}} \neq 0 \\ \frac{1}{m_k}, & \hat{\gamma_{k,j}} = 0\ \forall j = 1, \cdots, m_k \end{cases}, \quad k = 2, \cdots, 5$$

where $\hat{\gamma_{k,j}}$ are the estimates averaged over the twenty repetitions and m_k is the number of predictors included in RS k (compare Figure 1).

2.4. Risk Scores, Linear Model and Relative Importance in the Test Sample

The RS weights derived from the training sample are used to calculate the values of each RS k in the test sample as the weighted average of the respective predictors: $z_k = \sum_{j=1}^{m_k} w_{k,j}\ x_{k,j}, k = 2, \cdots, 5$. The risk score values are then scaled by their interquartile ranges: $\tilde{z}_k = \frac{z_k}{IQR(z_k)}, k = 2, \cdots, 5$.

Afterwards, the risk scores and the fixed single predictors are used as independent variables in a multiple linear regression model for each outcome (here: a lung function index):

$$y = \beta_0 + \underbrace{\sum_{j=1}^{4} \beta_{1,j}\ z_{1,j}}_{\text{Single predictors}} + \beta_2 \underbrace{\tilde{z}_2}_{\text{Genetic RS}} + \beta_3 \underbrace{\tilde{z}_3}_{\text{Obesity RS}} + \beta_4 \underbrace{\tilde{z}_4}_{\text{Smoking RS}} + \beta_5 \underbrace{\tilde{z}_5}_{\text{Air pollution RS}}$$

Finally, the relative importance of all independent variables in this linear model is calculated with the R function calc.relimp [42], where the relative importance of socio-economic status (SES) is assessed using the two binary dummy variables as one group. The relative importance metric Lindeman–Merenda–Gold is used, which decomposes the coefficient of determination of the model, R^2, by averaging sequential R^2s over orderings of regressors [18, chapter 4.7]). The relative contributions given by this metric sum to the overall R^2 [42].

All analysis steps are carried out for each bootstrap sample. The results presented in the following section are thus based on medians and percentiles across the bootstrap samples. The regression coefficients of the risk scores reflect the change in the outcome for an increase of one interquartile range of the respective RS and are used to determine significance of the association. Since the relative contributions of the risk scores are given as percentages, the corresponding bootstrap confidence intervals lie by definition above zero and cannot determine significance. All calculations and figures were done using R version 4.0.3 [43], except for Figure 1, which was produced using Microsoft Office Professional Plus 2019.

3. Results

3.1. Descriptive Analyses of the Outcomes and the Predictors

Descriptive statistics for all outcome and predictor variables are given in Table 1 and have been calculated for all 547 participants with available genetic and skin aging data and for 510 participants with available genetic and lung function data.

The two analysis samples differ only slightly in their characteristics. The women were on average 73 years old at the time of the second follow-up, were on average mildly overweight (mean BMI: 27 kg/m^2) and according to the mean MeDi score of 28.5 their nutritional habits moderately followed the Mediterranean diet. Only a few women were current or former smokers (2% and 16%, respectively), but about 33% were exposed to ETS at home and 42% at work.

3.2. Skin Aging Outcome

For the outcome of facial pigment spots all predictors combined explain 16.90% of the variance (median total R^2, Table 2). As can be seen from the regression coefficients in Figure 2, the genetic RS and sunbed use are associated with the formation of facial pigment spots (bootstrap medians, 95% percentile CIs and p-values: 0.29 [0.07, 0.50], p = 0.016;

0.15 [0.03, 0.29], $p = 0.016$) with the two highest median relative contributions of 4.15% and 2.11%.

The contributions of MeDi (1.20%) and air pollution RS (1.14%) seem interesting for future research according to their regression coefficients (0.11 [−0.01, 0.24], $p = 0.064$ and 0.16 [−0.04, 0.57], $p = 0.132$) whereby it is noticeable that stronger adherence to the Mediterranean diet seems to increase the formation of pigment spots.

Table 2. Relative Contributions (bootstrap median and 95% percentile confidence interval) of the predictors for pigment spots, as percentages, sorted according to the median.

Predictor	Median	95% Confidence Interval
Overall	*16.90%*	*8.75%, 28.04%*
Genetic RS	4.15%	0.25%, 11.88%
Sunbed use	2.11%	0.09%, 6.97%
Air pollution RS	1.20%	0.04%, 5.76%
MeDi	1.14%	0.04%, 5.11%
SES	0.70%	0.09%, 3.44%
Obesity RS	0.66%	0.04%, 4.59%
SPF	0.59%	0.04%, 3.33%
Skin type	0.58%	0.03%, 3.51%
Age	0.52%	0.04%, 3.37%
Solar radiation RS	0.45%	0.04%, 3.35%
HRT	0.29%	0.03%, 2.41%
Smoking RS	0.29%	0.02%, 2.74%

MeDi: Mediterranean diet; HRT: hormone replacement therapy; RS: risk score; SES: socio-economic status; SPF: sun protection factor.

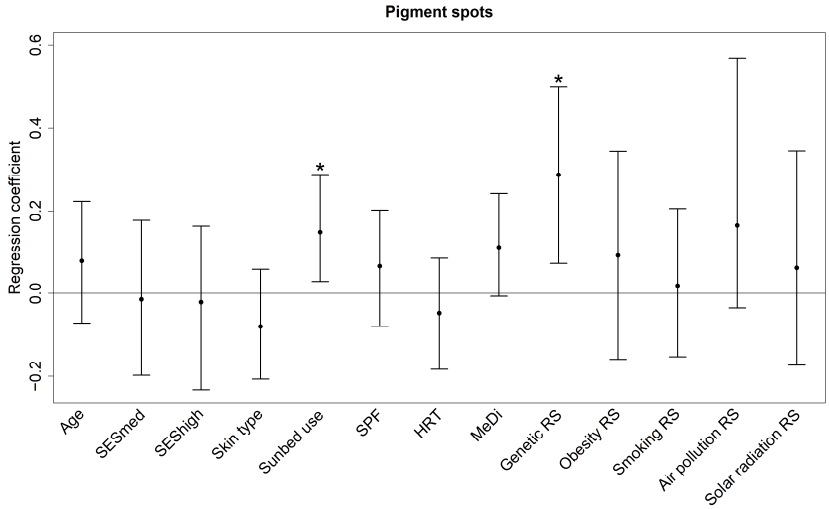

Figure 2. Regression coefficients (bootstrap medians and 95% percentile confidence intervals) of the predictors for pigment spots. The coefficients reflect the change in the z-score for one unit increase in the single predictors and one interquartile range increase in the risk scores. MeDi: Mediterranean diet; HRT: hormone replacement therapy; RS: risk score; SESmed: medium socio-economic status (reference: low socio-economic status); SEShigh: high socio-economic status (reference: low socio-economic status); SPF: sun protection factor; * $p < 0.05$.

3.3. Lung Function Outcomes

All single predictors and risk scores combined explain in median 22.32% of the variance in FEV_1 and 23.36% of the variance in FVC (see Tables 3 and 4). The genetic RS is associated with both lung function parameters according to its regression coefficients (see Figures 3 and 4; 0.43 [0.20, 0.63], $p = 0.01$ for FEV_1 and 0.39 [0.21, 0.59], $p = 0.01$ for FVC) with relative contributions of about 11%. The risk scores for genetics, smoking and obesity are among the top three relative contributors to both outcomes. For FEV_1 the obesity

(median relative contribution 2.55%) and smoking risk scores (median relative contribution 4.98%) are additionally associated (0.25 [0.04, 0.53], $p = 0.01$ and 0.14 [0.03, 0.33], $p = 0.00$), while for FVC it is only the obesity RS (median relative contribution 5.56%; coefficient 0.29 [0.11, 0.46], $p = 0.00$). The smoking risk score's bootstrap median coefficient for FVC is 0.09 with 95% percentile confidence interval [−0.01, 0.24], $p = 0.06$. It might seem as if the effects of obesity and smoking are beneficial due to the positive regression coefficients, but the risk scores' weights are mostly negative so that a larger RS refers to less obesity and less tobacco smoke exposure.

Table 3. Relative Contributions (bootstrap median and 95% percentile confidence interval) of the predictors for FEV_1, as percentages, sorted according to the median.

Predictor	Median	95% Confidence Interval
Overall	*22.32%*	*10.87%, 33.57%*
Genetic RS	10.99%	2.92%, 23.29%
Smoking RS	5.10%	1.15%, 11.87%
Obesity RS	2.41%	0.17%, 7.62%
SES	0.80%	0.13%, 3.48%
Height	0.45%	0.02%, 2.92%
Air pollution RS	0.36%	0.02%, 2.69%
Age	0.26%	0.04%, 1.85%

FEV_1: forced expiratory volume in 1 s; RS: risk score; SES: socio-economic status.

Table 4. Relative Contributions (bootstrap median and 95% percentile confidence interval) of the predictors for FVC, as percentages, sorted according to the median.

Predictor	Median	95% Confidence Interval
Overall	*23.36%*	*10.86%, 33.67%*
Genetic RS	11.69%	2.80%, 25.11%
Obesity RS	5.62%	1.21%, 12.50%
Smoking RS	1.83%	0.06%, 6.42%
SES	0.90%	0.13%, 3.71%
Air pollution RS	0.42%	0.03%, 3.40%
Age	0.26%	0.03%, 2.58%
Height	0.24%	0.02%, 1.96%

FVC: forced vital capacity; RS: risk score; SES: socio-economic status.

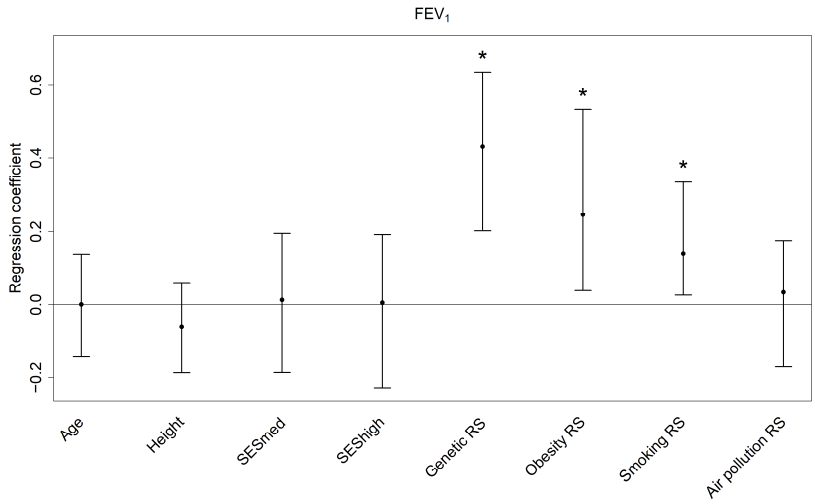

Figure 3. Regression coefficients (bootstrap median and 95% confidence interval) of the predictors for FEV_1. The coefficients reflect the change in the z-score for one unit increase in the single predictors and one interquartile range increase in the risk scores. FEV_1: forced expiratory volume in 1 s; RS: risk score; SESmed: medium socio-economic status (reference: low socio-economic status); SEShigh: high socio-economic status (reference: low socio-economic status); * $p < 0.05$.

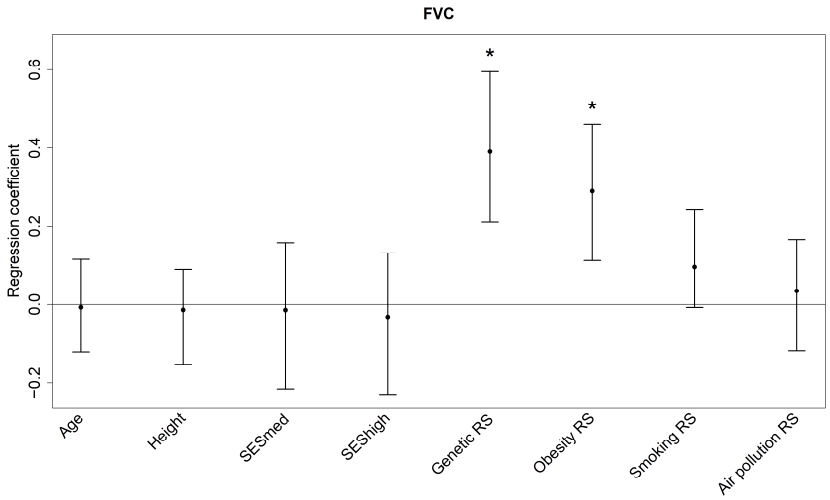

Figure 4. Regression coefficients (bootstrap median and 95% confidence interval) of the predictors for FVC. The coefficients reflect the change in the z-score for one unit increase in the single Predictors and one Interquartile Range Increase in the Risk Scores. FVC: forced vital capacity; RS: risk score; SESmed: medium socio-economic status (reference: low socio-economic status); SEShigh: high socio-economic status (reference: low socio-economic status); * $p < 0.05$.

4. Discussion

The proposed approach enables us to quantify the contributions of genetic factors and various categories of the exposome to a certain outcome in terms of percentages of explained variance, where each category has been assessed by several (correlated) variables and combined into one weighted RS.

In these examples, the highest contribution to all aging traits was achieved by the genetic risk scores comprising the considered SNPs. Apart from the genetic and the obesity RS and in parts the smoking RS, the environmental risk scores were not associated with the outcomes. That the lower limit of the confidence interval of the air pollution risk score's coefficient is very close to zero in the skin aging example is in line with previous analyses in the SALIA study, which found associations of air pollution with skin aging [44] when using single-pollutant models with no need to split off a training sample. One might have expected to see associations of the aging traits with age (at least for the skin aging outcome, where the z-score does not account for age). This is probably not only due to the small sample size, but also due to a small age range of the study participants.

The magnitudes and rankings of the estimated relative importance are quite similar compared between the two lung function parameters, while there are distinct differences when comparing the same predictors and risk scores between skin aging and lung function. The percentage of variance explained by the genetic RS in the skin aging outcome is only less than half of that in the lung function outcomes. In addition, the smoking risk score's relative contribution, for example, ranks very low for skin aging, while it has a top three ranking for lung function. Thus, awareness campaigns and other measures to reduce the number of smokers seem more important for reducing lung aging than for delaying skin aging in the population. Against skin aging it would be more effective to inform about the negative effects of sunbed use.

Overall, the models were able to explain between 17% and 23% of the different outcomes' variances, which is noticeable considering the complexity of the aging processes, the limited sample size and the fact that further components of the exposome such as stress or lack of sleep which were not collected in the SALIA study could not be incorporated. Though the presented example is far from a complete exposome analysis, this investigation shows that (i) in principle the proposed approach can quantify the extent to which each of the various categories of the exposome contribute, and that (ii) these relative contributions vary for different health traits and thus can be ranked.

The approach has some limitations. First, it relies on the calculation of relative contributions via shares of explained variance in a linear regression model and is as a consequence limited to linear regression. Generalized linear models such as logistic regression for binary traits are not applicable, since the concept of relative contribution is not (easily) generalizable. This direction is interesting for future research since binary health outcomes are very common. Second, splitting the available data set into training and test samples does not only reduce statistical power in the linear regression analysis, but also requires repeated execution of the fitting process (here: several bootstrap samples) to reduce the randomness of the partition and the results. This, however, complicates reporting of the results which have to be averaged. In particular, residual diagnostics are not easily applicable since they would have to be examined for each bootstrap sample. Yet, there is usually no alternative to internal weights for the risk scores, since external weights from published studies with several covariates belonging to the same risk factor category are typically not available precisely because they are often highly correlated. Third, combining several covariates in one (weighted) RS certainly leads to loss of information. An alternative would be to directly report the results of a regularized regression without the formation of risk scores. However, the aim of this study was to show to which extent each risk factor category contributes to the outcome. To the best of our knowledge, the concept of relative contribution is not (yet) applicable to regularized regression models. Such an extension is interesting for future research. In addition, this study is limited to a certain configuration of model parameters. For example, there might be choices other than $\alpha = 0$ (ridge regression) in the elastic net or other regularized regression or machine learning methods, which yield more appropriate RS weights for our purposes, but a comparison is beyond the scope of this work. Nevertheless, the presented application and results provide a proof of concept for the proposed methodology.

5. Conclusions

The combination of risk scores with a measure of relative contribution is suitable to assess the extent to which various categories of the exposome and genetic factors contribute to a certain health outcome, and the contributions can not only be compared between different outcomes, but also between, for example, different ethnic or age groups. In addition, the exposome's categories can be ranked according to their relative contribution. The proposed approach might thus have the potential to improve risk assessment relevant for human aging traits and beyond, i.e., common human diseases. In this regard it could be of interest not only to health scientists, but also to governmental institutions, because it might help to prioritize regulatory decisions limiting exposure to selected environmental factors and put them on a more solid scientific basis.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijerph192416746/s1, Supplementary File S1: Table S1: Scoring of the food frequencies for the Mediterranean diet score; Table S2: Single nucleotide polymorphisms used in the genetic risk score for the skin aging trait; Table S3: Single nucleotide polymorphisms used in the genetic risk score for the lung function traits; Figure S1: Correlation plot of the original variables used in the skin aging analysis (excluding the SNP variables); Figure S2: Correlation plot of the original variables used in the lung function analysis (excluding the SNP variables); Figure S3: Correlation plot of the risk scores and single predictors used in the skin aging analysis; Figure S4: Correlation plot of the risk scores and single predictors used in the analysis of FEV1; Figure S5: Correlation plot of the risk scores and single predictors used in the analysis of FVC. Refs. [23–40,45–47] are cited in the supplementary files.

Author Contributions: Conceptualization, J.K. and T.S.; Formal analysis, C.W.; Funding acquisition, T.S.; Methodology, C.W. and A.H.; Project administration, C.W. and T.S.; Software, C.W.; Supervision, J.K. and T.S.; Validation, C.W., A.H., J.K. and T.S.; Visualization, C.W.; Writing—original draft, C.W.; Writing—review & editing, C.W., A.H., J.K. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: The SALIA cohort study was supported by the Ministry of the Environment of the state North Rhine-Westphalia (Düsseldorf, Germany), the Federal Ministry of the Environment (Berlin, Germany), the German Federal Ministry of Education and Research (BMBF) as well as by grants HE-4510/2-1, KR 1938/3-1, LU 691/4-1 and SCHI 1358/3-1 from the Deutsche Forschungsgemeinschaft (DFG), VT 266.1 from the German Statutory Accident Insurance (DGUV) and grant agreement number 211250 from the European Community's Seventh Framework Program (FP7/2007-2011). Anke Hüls is supported by grant [NIEHS P30ES019776] from the HERCULES Center.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Medical Ethics Committee of the University of Bochum (approval number 2732, date of approval: 4 April 2006).

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Due to privacy laws in Germany the data of the cohort study are not available. The R code can be obtained by contacting the authors of the paper.

Acknowledgments: We thank all SALIA study members and staff involved in data collections, and also the respective funding bodies. SALIA study directorate: R. Dolgner, U. Krämer, U. Ranft, T. Schikowski. Scientific Team Baseline: A.W. Schlipköter, M.S. Islam, A. Brockhaus, H. Idel, R. Stiller-Winkler, W. Hadnagy, T. Eikmann. Scientific Team Follow-up: D. Sugiri, A. Hüls, B. Pesch, A. Hartwig, H. Käfferlein, V. Harth, T. Brüning, T. Weiss. Study Nurses: G. Seitner-Sorge, V. Jäger, G. Petczelies, I. Podolski, T. Hering, M.Goseberg. Administrative Team: B. Schulten, S. Stolz. During the last decades many scientists, study nurses, and laboratories were involved in conducting the study. We are most grateful to all the women from the Ruhr area and Borken who participated in the study over the course of decades and the local health departments for organizing the study. We are most indebted to D. Sugiri, who performed modelling of air pollution with the ESCAPE LUR model, and who assigned residential levels of UV radiation, air pollution and traffic indicators to the addresses of the SALIA study participants. This work has been supported in part by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation—Project Number 427806116).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Wild, C.P. Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiol. Biomark. Prev.* **2005**, *14*, 1847–1850. [CrossRef] [PubMed]
2. Kim, K.-N.; Hong, Y.-C. The exposome and the future of epidemiology: A vision and prospect. *Environ. Health Toxicol.* **2017**, *32*, e2017009. [CrossRef]
3. Siroux, V.; Agier, L.; Slama, R. The exposome concept: A challenge and a potential driver for environmental health research. *Eur. Respir. Rev.* **2016**, *25*, 124–129. [CrossRef] [PubMed]
4. Guillien, A.; Cadiou, S.; Slama, R.; Siroux, V. The Exposome Approach to Decipher the Role of Multiple Environmental and Lifestyle Determinants in Asthma. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1138. [CrossRef] [PubMed]
5. Patel, C.J.; Bhattacharya, J.; Butte, A.J. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. *PLoS ONE* **2010**, *5*, e10746. [CrossRef]
6. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
7. Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
8. Sinisi, S.E.; van der Laan, M.J. Deletion/substitution/addition algorithm in learning with applications in genomics. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 18. [CrossRef] [PubMed]
9. Carrico, C.; Gennings, C.; Wheeler, D.C.; Factor-Litvak, P. Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting. *J. Agric. Biol. Environ. Stat.* **2015**, *20*, 100–120. [CrossRef] [PubMed]
10. Keil, A.P.; Buckley, J.P.; O'Brien, K.M.; Ferguson, K.K.; Zhao, S.; White, A.J. A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures. *Environ. Health Perspect.* **2020**, *128*, 047004. [CrossRef]
11. Bobb, J.F.; Valeri, L.; Claus Henn, B.; Christiani, D.C.; Wright, R.O.; Mazumdar, M.; Godleski, J.J.; Coull, B.A. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **2015**, *16*, 493–508. [CrossRef] [PubMed]
12. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **2013**, *9*, e1003348. [CrossRef]

13. Forgetta, V.; Keller-Baruch, J.; Forest, M.; Durand, A.; Bhatnagar, S.; Kemp, J.P.; Nethander, M.; Evans, D.; Morris, J.A.; Kiel, D.P.; et al. Development of a polygenic risk score to improve screening for fracture risk: A genetic risk prediction study. *PLoS Med.* **2020**, *17*, e1003152. [CrossRef] [PubMed]
14. Mavaddat, N.; Michailidou, K.; Dennis, J.; Lush, M.; Fachal, L.; Lee, A.; Tyrer, J.P.; Chen, T.-H.; Wang, Q.; Bolla, M.K.; et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **2019**, *104*, 21–34. [CrossRef] [PubMed]
15. Hüls, A.; Krämer, U.; Carlsten, C.; Schikowski, T.; Ickstadt, K.; Schwender, H. Comparison of weighting approaches for genetic risk scores in gene-environment interaction studies. *BMC Genet.* **2017**, *18*, 115. [CrossRef] [PubMed]
16. Lin, W.-Y.; Lin, Y.-S.; Chan, C.-C.; Liu, Y.-L.; Tsai, S.-J.; Kuo, P.-H. Using Genetic Risk Score Approaches to Infer Whether an Environmental Factor Attenuates or Exacerbates the Adverse Influence of a Candidate Gene. *Front. Genet.* **2020**, *11*, 331. [CrossRef]
17. Lin, W.-Y.; Huang, C.-C.; Liu, Y.-L.; Tsai, S.-J.; Kuo, P.-H. Polygenic approaches to detect gene–environment interactions when external information is unavailable. *Brief Bioinform.* **2019**, *20*, 2236–2252. [CrossRef]
18. Lindeman, R.H.; Merenda, P.F.; Gold, R.Z. *Introduction to Bivariate and Multivariate Analysis*; Scott Foresman & Co.: Glenview, IL, USA, 1980.
19. Krutmann, J.; Schikowski, T.; Morita, A.; Berneburg, M. Environmentally-Induced (Extrinsic) Skin Aging: Exposomal Factors and Underlying Mechanisms. *J. Invest. Dermatol.* **2021**, *141*, 1096–1103. [CrossRef]
20. Wheelock, C.E.; Rappaport, S.M. The role of gene–environment interactions in lung disease: The urgent need for the exposome. *Eur. Respir. J.* **2020**, *55*, 1902064. [CrossRef]
21. Götschi, T.; Heinrich, J.; Sunyer, J.; Künzli, N. Long-term effects of ambient air pollution on lung function: A review. *Epidemiology* **2008**, *19*, 690–701. [CrossRef]
22. U.S. Department of Health and Human Services. *The Health Consequences of Smoking: A Report of the Surgeon General*; U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health: Atlanta, GA, USA, 2004.
23. Schikowski, T.; Sugiri, D.; Ranft, U.; Gehring, U.; Heinrich, J.; Wichmann, H.E.; Krämer, U. Long-term air pollution exposure and living close to busy roads are associated with COPD in women. *Respir. Res.* **2005**, *6*, 152–162. [CrossRef] [PubMed]
24. Schikowski, T.; Sugiri, D.; Ranft, U.; Gehring, U.; Heinrich, J.; Wichmann, H.E.; Krämer, U. Does respiratory health contribute to the effects of long-term air pollution exposure on cardiovascular mortality? *Respir. Res.* **2007**, *8*, 20. [CrossRef] [PubMed]
25. Schikowski, T.; Vossoughi, M.; Vierkötter, A.; Schulte, T.; Teichert, T.; Sugiri, D.; Fehsel, K.; Tzivian, L.; Bae, I.S.; Ranft, U.; et al. Association of air pollution with cognitive functions and its modification by APOE gene variants in elderly women. *Environ. Res.* **2015**, *142*, 10–16. [CrossRef]
26. Miller, M.R.; Hankinson, J.; Brusasco, V.; Burgos, F.; Casaburi, R.; Coates, A.; Crapo, R.; Enright, P.; van der Grinten, C.P.; Gustafsson, P.; et al. Standardisation of spirometry. *Eur. Respir. J.* **2005**, *26*, 319–338. [CrossRef] [PubMed]
27. Quanjer, P.H.; Stanojevic, S.; Cole, T.J.; Baur, X.; Hall, G.L.; Culver, B.H.; Enright, P.L.; Hankinson, J.L.; Ip, M.S.M.; Zheng, J.; et al. Multi-ethnic reference values for spirometry for the 3–95-yr age range: The global lung function 2012 equations. *Eur. Respir. J.* **2012**, *40*, 1324–1343. [CrossRef] [PubMed]
28. Tschachler, E.; Morizot, F. Ethnic Differences in Skin Aging. In *Skin Aging*; Krutmann, J., Gilchrest, B.A., Eds.; Springer GmbH: Berlin, Germany, 2006.
29. Vierkötter, A.; Ranft, U.; Krämer, U.; Sugiri, D.; Reimann, V.; Krutmann, J. The SCINEXA: A novel, validated score to simultaneously assess and differentiate between intrinsic and extrinsic skin ageing. *J. Dermatol. Sci.* **2009**, *53*, 207–211. [CrossRef] [PubMed]
30. Das, S.; Forer, L.; Schönherr, S. Next-generation genotype imputation service and methods. *Nat. Genet.* **2016**, *48*, 1284–1287. [CrossRef]
31. Endo, C.; Johnson, T.A.; Morino, R.; Nakazono, K.; Kamitsuji, S.; Akita, M.; Kawajiri, M.; Yamasaki, T.; Kami, A.; Hoshi, Y.; et al. Genome-wide association study in Japanese females identifies fifteen novel skin-related trait associations. *Sci. Rep.* **2018**, *8*, 8974. [CrossRef]
32. Jacobs, L.C.; Hamer, M.A.; Gunn, D.A.; Deelen, J.; Lall, J.S.; van Heemst, D.; Uh, H.W.; Hofman, A.; Uitterlinden, A.G.; Griffiths, C.E.M.; et al. A Genome-Wide Association Study Identifies the Skin Color Genes IRF4, MC1R, ASIP, and BNC2 Influencing Facial Pigmented Spots. *J. Invest. Dermatol.* **2015**, *135*, 1735–1742. [CrossRef]
33. Laville, V.; Clerc, S.L.; Ezzedine, K.; Jdid, R.; Taing, L.; Labib, T.; Coulonges, C.; Ulveling, D.; Carpentier, W.; Galan, P.; et al. A genome-wide association study in Caucasian women suggests the involvement of HLA genes in the severity of facial solar lentigines. *Pigment Cell Melanoma Res.* **2016**, *29*, 550–558. [CrossRef]
34. Liu, F.; Hamer, M.A.; Deelen, J.; Lall, J.S.; Jacobs, L.; van Heemst, D.; Murray, P.G.; Wollstein, A.; de Craen, A.J.; Uh, H.W.; et al. The MC1R Gene and Youthful Looks. *Curr. Biol.* **2016**, *26*, 1213–1220. [CrossRef] [PubMed]
35. Shin, J.-G.; Leem, S.; Kim, B.; Kim, Y.; Lee, S.-G.; Song, H.J.; Seo, J.Y.; Park, S.G.; Won, H.-H.; Kang, N.G. GWAS Analysis of 17,019 Korean Women Identifies the Variants Associated with Facial Pigmented Spots. *J. Invest. Dermatol.* **2021**, *141*, 555–562. [CrossRef]
36. Shrine, N.; Guyatt, A.L.; Erzurumluoglu, A.M.; Jackson, V.E.; Hobbs, B.D.; Melbourne, C.A.; Batini, C.; Fawcett, K.A.; Song, K.; Sakornsakolpat, P.; et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **2019**, *51*, 481–493. [CrossRef] [PubMed]
37. Adam, M.; Schikowski, T.; Carsin, A.E.; Cai, Y.; Jacquemin, B.; Sanchez, M.; Vierkötter, A.; Marcon, A.; Keidel, D.; Sugiri, D.; et al. Adult lung function and long-term air pollution exposure. ESCAPE: A multicentre cohort study and meta-analysis. *Eur. Respir. J.* **2014**, *45*, 38–50. [CrossRef]

38. Beelen, R.; Hoek, G.; Vienneau, D.; Eeftens, M.; Dimakopoulou, K.; Pedeli, X.; Tsai, M.-Y.; Künzli, N.; Schikowski, T.; Marcon, A.; et al. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe—The ESCAPE project. *Atmos. Environ.* **2013**, *72*, 10–23. [CrossRef]
39. Eeftens, M.; Beelen, R.; de Hoogh, K.; Bellander, T.; Cesaroni, G.; Cirach, M.; Declercq, C.; Dedele, A.; Dons, E.; de Nazelle, A.; et al. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas;results of the ESCAPE project. *Environ. Sci. Technol.* **2012**, *46*, 11195–11205. [CrossRef]
40. Hüls, A.; Sugiri, D.; Fuks, K.; Krutmann, J.; Schikowski, T. Lentigine Formation in Caucasian Women—Interaction between Particulate Matter and Solar UVR. *J. Invest. Dermatol.* **2019**, *139*, 974–976. [CrossRef]
41. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
42. Grömping, U. Relative Importance for Linear Regression in R: The Package relaimpo. *J. Stat. Softw.* **2006**, *17*, 1–27. [CrossRef]
43. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
44. Vierkötter, A.; Schikowski, T.; Ranft, U.; Sugiri, D.; Matsui, M.; Krämer, U.; Krutmann, J. Airborne particle exposure and extrinsic skin aging. *J. Invest. Dermatol.* **2010**, *130*, 2719–2726. [CrossRef]
45. Panagiotakos, D.B.; Pitsavos, C.; Arvaniti, F.; Stefanadis, C. Adherence to the Mediterranean food pattern predicts the prevalence of hypertension, hypercholesterolemia, diabetes and obesity, among healthy adults; the accuracy of the MedDietScore. *Prev. Med.* **2007**, *44*, 335–340. [CrossRef] [PubMed]
46. Schwender, H. "scrime": Analysis of High-Dimensional Categorical Data Such as SNP Data. R Package Version 1.3.5. 2018. Available online: https://cran.r-project.org/web/packages/scrime/scrime.pdf (accessed on 8 November 2022).
47. Wei, T.; Simko, V. R Package "corrplot": Visualization of a Correlation Matrix. R Package Version 0.84. 2017. Available online: https://cran.r-project.org/web/packages/corrplot/corrplot.pdf (accessed on 8 November 2022).