Supplementary Methods

Title: Occupational exposure to pesticides is associated with differential DNA methylation

Authors: Diana A. van der Plaat[1,2], Kim de Jong[1,2], Maaike de vries[1,2], Cleo C. van Diemen[3], Ivana Nedeljkovic[4], Najaf Amin[4], BIOS consortium, Hans Kromhout[5], Roel Vermeulen[5], Dirkje S. Postma[6,2], Cornelia M. van Duijn[3], H. Marike Boezen[1,2], Judith M. Vonk[1,2]

Author affiliation: [1]University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, the Netherlands. [2]University of Groningen, University Medical Center Groningen, Groningen Research Institute for Asthma and COPD (GRIAC), Groningen, the Netherlands. [3]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands. [4]Erasmus Medical Center, Department of Epidemiology, Rotterdam, the Netherlands. [5]Utrecht University, Institute for Risk Assessment Sciences, Division Environmental Epidemiology, Utrecht, the Netherlands. [6]University of Groningen, University Medical Center Groningen, Department of Pulmonary Diseases, Groningen, the Netherlands.

LifeLines Cohort Study

LifeLines is a multi-disciplinary prospective population-based cohort study examining in a unique three-generation design the health and health-related behaviors of 167,729 persons living in the North of the Netherlands.[1] It employs a broad range of investigative procedures in assessing the biomedical, socio-demographic, behavioral, physical and psychological factors which contribute to the health and disease of the general population, with a special focus on multi-morbidity and complex genetics.

Quality control 450K arrays

Quality-control (QC) steps were performed using the Minfi package in R, and included the removal of samples with >1% of all probes having a detection p-value >0.01, and samples with an incorrect sex or SNP prediction.[2]We removed single probes with a detection p-value >0.01, sex chromosome probes, cross-reactive probes, probes measuring SNPs, and probes where the CpG itself or the single base extension (SBE) site is a SNP.[3] The data were normalized using DASEN implemented in the wateRmelon package in R.[4]

Principal component (PC) analysis

To estimate possible batch effects, a principal component (PC) analysis was performed using the control probes included on the 450K chip.[5] In total, 220 control probes are included on each chip, related to bisulfite conversion, normalization, staining, extension, hybridization, target removal, specificity and non-polymorphic. We selected all PCs that explained at least 1% of the variance. Therefore 7 PCs were selected and these captured in total 95.5% of the technical variance (Table E1). Next, we correlated the PCs with possible known batch effect variables. The headmap shows that variation in the position on the chip, date of bisulfite treatment, chip number and sample plate were correlated with these PCs, especially in the first PC (Figure E1 and E2).

Next, we also calculated the PCs of the 1,561 subjects and 420,938 CpG sites available after QC (excluding the control probes). To assess if the PCs that explained >1% of the variance were associated with pesticide exposure we performed a linear regression with the low and high pesticide exposure variables as independent variables. The results show that none of the PCs was significantly associated with low or high pesticide exposure (table E2, model 1). Also, if we exclude the potential confounders sex, age, smoking and white blood cell counts from the principal component analysis, we do not find a significant association between the PCs and pesticide exposure (Table E2, model 2). This could indicate that exposure to pesticides is associated with DNA methylation variation at specific sites in the genome.

Table E1. Principal component analysis performed in the control probes.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Proportion Of Variance	70.8%	9.9%	6.2%	3.7%	2.2%	1.5%	1.1%
Cumulative Proportion	70.8%	80.7%	86.9%	90.6%	92.8%	94.3%	95.5%

Table E2. Association of PCs with pesticide exposure (PCs were based on CpGs after QC).

	Model 1							***Model 2*****						
		Low exposure			*High exposure*				*Low exposure*			*High exposure*		
PC	**PoV***	**B**	**SE**	**P**	**B**	**SE**	**P**	**PoV***	**B**	**SE**	**P**	**B**	**SE**	**P**
PC1	12.3%	-1.39	0.72	0.052	-1.05	0.94	0.264	9.1%	-0.60	0.59	0.306	-0.34	0.77	0.656
PC2	5.3%	-0.44	0.38	0.246	-0.46	0.50	0.357	5.4%	-0.45	0.38	0.243	-0.60	0.50	0.232
PC3	1.9%	0.02	0.27	0.942	-0.21	0.36	0.555	2.0%	0.00	0.28	0.996	-0.46	0.37	0.215
PC4	1.8%	0.09	0.27	0.754	-0.39	0.36	0.280	1.5%	-0.24	0.23	0.301	-0.22	0.30	0.465
PC5	1.3%	-0.13	0.23	0.571	0.04	0.31	0.902	1.3%	0.00	0.22	0.999	0.23	0.28	0.418
PC6	1.2%	-0.14	0.21	0.494	-0.19	0.27	0.478	1.2%	0.06	0.22	0.785	-0.25	0.29	0.386

* PoV = Proportion of Variance explained by PC. **The potential confounders sex, age, smoking and white blood cell counts were excluded from the principal component analysis.

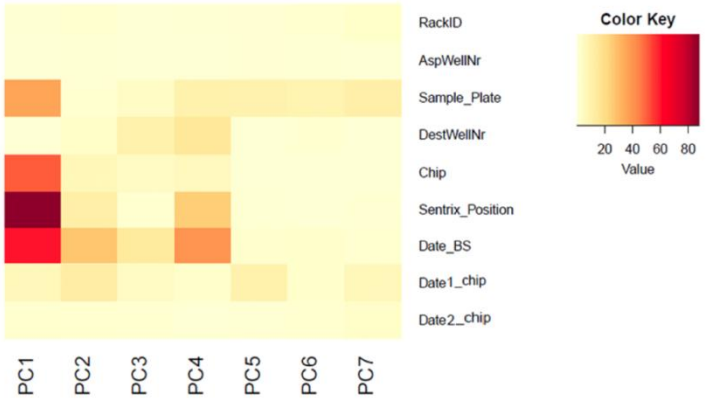

Figure E1. Heatmap of the correlation between the seven principal components and possible known batch effects. The more red the color is, the higher the correlation between variables.

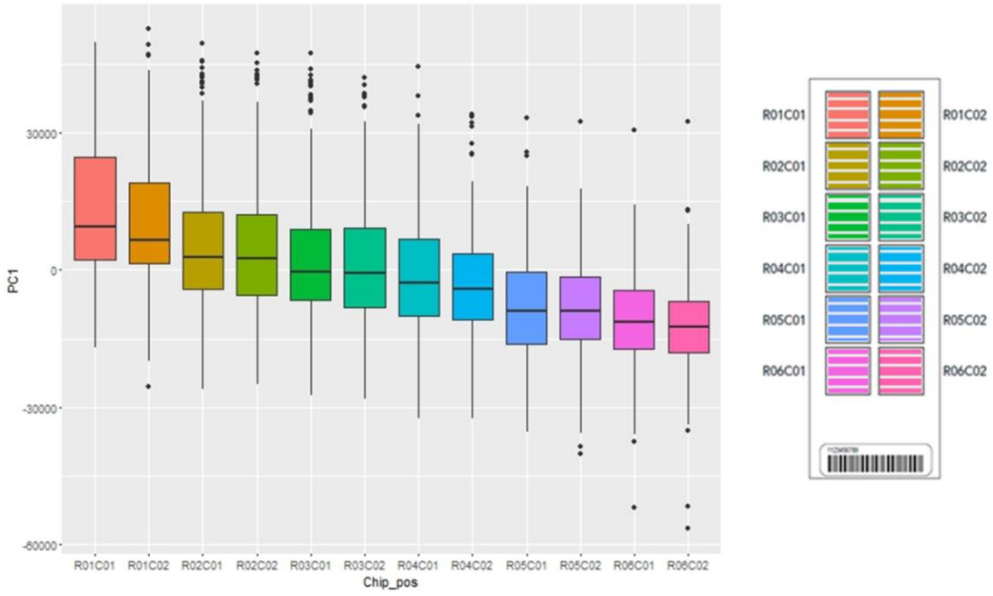

Figure E2. Plot of principal component 1 versus the position on the chip.

Additional analyses

Differential blood counts

In the current study we had differential blood counts available obtained using standard laboratory techniques for lymphocytes, monocytes, and eosinophilic, neutrophilic, and basophilic granulocytes. In addition, we used the *estimateCellCounts* function in the R package *Minfi* to estimates the relative proportion of pure cell types based on the Houseman method.[6,7] We also calculated the Neutrophil lymphocyte ratio (NLR) for the measured cell counts by dividing the absolute neutrophil count by the absolute lymphocyte count, because the NLR is a prognostic biomarker for many different diseases.[8] Subsequently, to assess whether cell composition, either measured or predicted, associated with pesticide exposure, we performed linear regression using the measured and predicted differential blood counts or the NLR as dependent variable and pesticide exposure (low and high exposure dummies) as independent variables. Looking at the results (table E3), high pesticide exposure is not significantly associated with any of the blood cell types, measured or predicted. Percentage of measured neutrophils and lymphocytes, and predicted CD8T, natural killer cells and granulocytes are nominally associated with low pesticide exposure, but none surviving adjustment for multiple testing. The NLR is not significantly associated with low or high pesticide exposure in our dataset. Although it is known that pesticides can alter the immune function [9], we do not see clear evidence in our study.

Table E3.Association between blood cell counts and pesticide exposure.

	Low pesticide exposure			*High pesticide exposure*		
	B	**SE**	**P**	**B**	**SE**	**P**
Measured cell counts						
Basophilic Granulocytes (%)	-0.01	0.03	0.814	0.08	0.04	0.068
Eosinophil Granulocytes (%)	-0.22	0.19	0.254	0.00	0.25	0.987
Neutrophil Granulocytes (%)	2.01	0.83	**0.016**	1.10	1.09	0.312
Lymphocytes (%)	-1.92	0.75	**0.011**	-1.01	0.99	0.308
Mononuclear cells (%)	0.14	0.19	0.458	-0.18	0.24	0.466
NLR (ratio GR / LY)	0.16	0.09	0.063	0.06	0.12	0.630
Predicted cell counts*						
CD8T cells (%)	-0.01	0.00	**0.045**	-0.01	0.00	0.173
CD4T cells (%)	-0.01	0.01	0.383	-0.01	0.01	0.533
Natural killer cells (%)	-0.01	0.01	**0.037**	-0.01	0.01	0.348
B-cells (%)	0.00	0.00	0.100	0.01	0.00	0.173
Mononuclear cells (%)	0.00	0.00	0.611	0.00	0.00	0.444
Granulocytes (%)	0.03	0.01	**0.043**	0.01	0.02	0.520

All models were adjusted for sex, age, smoking and pack-years (N=1,561).

*Predicted using the *estimateCellCounts* function in the R package *Minfi* which calculates predicted white blood cell counts based on Houseman et al.[6]

KEGG pathway and GO term enrichment analyses

We have performed KEGG pathway (Kyoto Encyclopedia of Genes and Genomes, n=321) and GO term (Gene Ontology terms, n=21,681) enrichment analyses using the *gometh* function in the *missMethyl* package in R.[10] This function takes the significant CpG sites, maps the CpG sites to Entrez Gene IDs, and tests for GO term or KEGG pathway enrichment using a hypergeometric test, taking into account the number of CpG sites per gene on the 450K array. We used all CpGs with a p-value $<10^{-4}$ for the analysis in the complete cohort and in the stratified analyses based on gender, smoking or airway obstruction separately. None of the KEGG pathways or GO terms had an FDR-adjusted p-value <0.05 in all analysis, except for five GO terms in females (Table E4). The results of the GO terms with a p-value <0.001 are shown in table E5 and of the KEGG pathways with a p-value <0.05 in table E6. The significant GO terms in females are: homophilic cell adhesion via plasma membrane adhesion molecules, cell-cell signaling, cell-cell adhesion via plasma-membrane adhesion molecules, calcium ion binding, and plasma membrane part.

Table E4. Results of GO term and KEGG pathway enrichment analysis.

		N GO terms*			N KEGG pathways*	
	N CpGs	**<FDR**	**<0.001†**	**<0.05**	**<FDR**	**<0.05‡**
Complete cohort	49	0	12	367	0	4
Males	88	0	1	331	0	3
Females	67	5	23	369	0	11
Never-smokers	81	0	20	314	0	2
Current-smokers	62	0	0	284	0	0
Subjects without airway obstruction	62	0	12	472	0	2
Subjects with airway obstruction	124	0	7	173	0	7

* Number of GO terms or KEGG pathways with a p-value below the threshold (FDR, 0.001 or 0.05)

† All GO terms with a p-value <0.001 are shown in table E4.

‡ All KEGG pathways with a p-value <0.05 are shown in table E5.

Table E5. GO terms with a p-value <0.001.

Analysis	GO	Term	Ont	N	DE	P.DE	FDR
Complete_cohort	GO:1903123	regulation of thioredoxin peroxidase activity*	BP	1	1	5.61E-04	1
Complete_cohort	GO:1903124	negative regulation of thioredoxin peroxidase activity*	BP	1	1	5.61E-04	1
Complete_cohort	GO:1903125	negative regulation of thioredoxin peroxidase activity by peptidyl-threonine phosphorylation*	BP	1	1	5.61E-04	1
Complete_cohort	GO:1904887	Wnt signalosome assembly*	BP	1	1	5.61E-04	1
Complete_cohort	GO:1905289	regulation of CAMKK-AMPK signaling cascade*	BP	1	1	5.61E-04	1
Complete_cohort	GO:2000469	negative regulation of peroxidase activity*	BP	1	1	5.61E-04	1
Complete_cohort	GO:0032473	cytoplasmic side of mitochondrial outer membrane*	CC	1	1	5.61E-04	1
Complete_cohort	GO:0099400	caveola neck*	CC	1	1	5.61E-04	1
Complete_cohort	GO:0034211	GTP-dependent protein kinase activity*	MF	1	1	5.61E-04	1
Complete_cohort	GO:0036479	peroxidase inhibitor activity*	MF	1	1	5.61E-04	1
Complete_cohort	GO:1904713	beta-catenin destruction complex binding*	MF	1	1	5.61E-04	1
Complete_cohort	GO:0005770	late endosome	CC	201	4	8.63E-04	1
Females	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	BP	150	11	1.76E-10	3.81E-06
Females	GO:0007267	cell-cell signaling	BP	1484	23	3.09E-09	2.45E-05
Females	GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	BP	221	11	3.39E-09	2.45E-05
Females	GO:0005509	calcium ion binding	MF	664	13	1.10E-06	5.96E-03
Females	GO:0044459	plasma membrane part	CC	2391	22	1.00E-05	0.044
Females	GO:0098609	cell-cell adhesion	BP	749	12	2.50E-05	0.090
Females	GO:0007155	cell adhesion	BP	1285	15	7.42E-05	0.214
Females	GO:0022610	biological adhesion	BP	1291	15	7.90E-05	0.214
Females	GO:0044700	single organism signaling	BP	5715	33	2.14E-04	0.395
Females	GO:0023052	signaling	BP	5722	33	2.22E-04	0.395
Females	GO:0007154	cell communication	BP	5741	33	2.24E-04	0.395
Females	GO:0048856	anatomical structure development	BP	5154	32	2.41E-04	0.395
Females	GO:0042734	presynaptic membrane	CC	66	4	2.64E-04	0.395
Females	GO:0005887	integral component of plasma membrane	CC	1451	14	2.67E-04	0.395
Females	GO:0007399	nervous system development	BP	2093	20	2.74E-04	0.395
Females	GO:0005886	plasma membrane	CC	4718	28	3.50E-04	0.474
Females	GO:0031226	intrinsic component of plasma membrane	CC	1511	14	4.28E-04	0.546
Females	GO:0046872	metal ion binding	MF	3825	25	4.89E-04	0.589

Females	GO:0071944	cell periphery	CC	4813	28	5.18E-04	0.591
Females	GO:0033981	D-dopachrome decarboxylase activity	MF	1	1	6.07E-04	0.616
Females	GO:0044425	membrane part	CC	5997	31	6.19E-04	0.616
Females	GO:0043169	cation binding	MF	3905	25	6.25E-04	0.616
Females	GO:0032502	developmental process	BP	5521	32	8.11E-04	0.765
Males	GO:0001938	positive regulation of endothelial cell proliferation	BP	68	4	3.85E-04	1
Never-smokers	GO:0005525	GTP binding	MF	341	7	1.15E-04	0.539
Never-smokers	GO:0032550	purine ribonucleoside binding	MF	343	7	1.18E-04	0.539
Never-smokers	GO:0001883	purine nucleoside binding	MF	346	7	1.24E-04	0.539
Never-smokers	GO:0032549	ribonucleoside binding	MF	347	7	1.27E-04	0.539
Never-smokers	GO:0001882	nucleoside binding	MF	353	7	1.36E-04	0.539
Never-smokers	GO:0032561	guanyl ribonucleotide binding	MF	358	7	1.72E-04	0.539
Never-smokers	GO:0019001	guanyl nucleotide binding	MF	359	7	1.74E-04	0.539
Never-smokers	GO:0033267	axon part	CC	168	6	2.57E-04	0.663
Never-smokers	GO:1903123	regulation of thioredoxin peroxidase activity*	BP	1	1	5.81E-04	0.663
Never-smokers	GO:1903124	negative regulation of thioredoxin peroxidase activity*	BP	1	1	5.81E-04	0.663
Never-smokers	GO:1903125	negative regulation of thioredoxin peroxidase activity by peptidyl-threonine phosphorylation*	BP	1	1	5.81E-04	0.663
Never-smokers	GO:1904887	Wnt signalosome assembly*	BP	1	1	5.81E-04	0.663
Never-smokers	GO:1905289	regulation of CAMKK-AMPK signaling cascade*	BP	1	1	5.81E-04	0.663
Never-smokers	GO:2000469	negative regulation of peroxidase activity*	BP	1	1	5.81E-04	0.663
Never-smokers	GO:0032473	cytoplasmic side of mitochondrial outer membrane*	CC	1	1	5.81E-04	0.663
Never-smokers	GO:0099400	caveola neck*	CC	1	1	5.81E-04	0.663
Never-smokers	GO:0034211	GTP-dependent protein kinase activity*	MF	1	1	5.81E-04	0.663
Never-smokers	GO:0036479	peroxidase inhibitor activity*	MF	1	1	5.81E-04	0.663
Never-smokers	GO:1904713	beta-catenin destruction complex binding*	MF	1	1	5.81E-04	0.663
Never-smokers	GO:0034260	negative regulation of GTPase activity	BP	37	3	8.36E-04	0.906
No_obstruction	GO:0048741	skeletal muscle fiber development	BP	30	3	1.06E-04	0.856
No_obstruction	GO:0007519	skeletal muscle tissue development	BP	145	5	1.18E-04	0.856
No_obstruction	GO:0060538	skeletal muscle organ development	BP	152	5	1.48E-04	0.856
No_obstruction	GO:0014904	myotube cell development	BP	33	3	1.58E-04	0.856
No_obstruction	GO:0048747	muscle fiber development	BP	50	3	4.71E-04	1
No_obstruction	GO:0002066	columnar/cuboidal epithelial cell development	BP	52	3	6.83E-04	1

No_obstruction	GO:0014706	striated muscle tissue development	BP	327	6	6.84E-04	1
No_obstruction	GO:0032703	negative regulation of interleukin-2 production	BP	16	2	7.46E-04	1
No_obstruction	GO:0007517	muscle organ development	BP	340	6	8.10E-04	1
No_obstruction	GO:0060537	muscle tissue development	BP	341	6	8.38E-04	1
No_obstruction	GO:0045927	positive regulation of growth	BP	227	5	8.53E-04	1
No_obstruction	GO:0071407	cellular response to organic cyclic compound	BP	518	7	9.88E-04	1
Obstruction	GO:0005604	basement membrane	CC	87	6	3.94E-05	0.447
Obstruction	GO:0032836	glomerular basement membrane development	BP	10	3	4.13E-05	0.447
Obstruction	GO:0044420	extracellular matrix component	CC	113	6	1.59E-04	1
Obstruction	GO:0005587	collagen type IV trimer	CC	4	2	4.58E-04	1
Obstruction	GO:0098642	network-forming collagen trimer	CC	5	2	6.22E-04	1
Obstruction	GO:0098645	collagen network	CC	5	2	6.22E-04	1
Obstruction	GO:0098651	basement membrane collagen trimer	CC	6	2	7.83E-04	1

** Identified in both the complete cohort and never-smokers.*

Ont = ontology that the GO term belongs to: "BP" - biological process, "CC" - cellular component, "MF" - molecular function.

N = number of genes in the GO term. DE = number of genes that are differentially methylated. P.DE = p-value for over-representation of the GO term.

FDR = False discovery rate.

Table E6.KEGG pathways with a p-value <0.05.

Analysis	KEGG	Pathway	N	DE	P.DE	FDR
Complete_cohort	path:hsa00270	Cysteine and methionine metabolism	43	2	0.006	1
Complete_cohort	path:hsa04014	Ras signaling pathway	226	3	0.034	1
Complete_cohort	path:hsa04713	Circadian entrainment	95	2	0.045	1
Complete_cohort	path:hsa04145	Phagosome	141	2	0.049	1
Males	path:hsa04015	Rap1 signaling pathway*	206	4	0.020	1
Males	path:hsa05014	Amyotrophic lateral sclerosis (ALS)	51	2	0.023	1
Males	path:hsa05230	Central carbon metabolism in cancer	62	2	0.047	1
Females	path:hsa05205	Proteoglycans in cancer	198	5	0.001	0.282
Females	path:hsa05012	Parkinson's disease	122	3	0.004	0.705
Females	path:hsa05206	MicroRNAs in cancer	278	4	0.008	0.813
Females	path:hsa05219	Bladder cancer	40	2	0.011	0.868
Females	path:hsa04721	Synaptic vesicle cycle	63	2	0.020	1
Females	path:hsa05217	Basal cell carcinoma	63	2	0.029	1
Females	path:hsa04024	cAMP signaling pathway	192	3	0.031	1
Females	path:hsa04971	Gastric acid secretion	74	2	0.033	1
Females	path:hsa04512	ECM-receptor interaction**	80	2	0.041	1
Females	path:hsa04670	Leukocyte transendothelial migration	106	2	0.046	1
Females	path:hsa04015	Rap1 signaling pathway*	206	3	0.047	1
Never-smokers	path:hsa04910	Insulin signaling pathway	132	3	0.020	1
Never-smokers	path:hsa04930	Type II diabetes mellitus	45	2	0.030	1
No_obstruction	path:hsa04550	Signaling pathways regulating pluripotency of stem cells	135	3	0.020	1
No_obstruction	path:hsa04730	Long-term depression	58	2	0.025	1
Obstruction	path:hsa00260	Glycine, serine and threonine metabolism	35	2	0.005	0.832
Obstruction	path:hsa04510	Focal adhesion	191	5	0.007	0.832
Obstruction	path:hsa04974	Protein digestion and absorption	78	3	0.009	0.832
Obstruction	path:hsa05146	Amoebiasis	92	3	0.010	0.832
Obstruction	path:hsa04512	ECM-receptor interaction**	80	3	0.014	0.908
Obstruction	path:hsa04926	Relaxin signaling pathway	125	3	0.035	1
Obstruction	path:hsa04151	PI3K-Akt signaling pathway	326	5	0.037	1

** Identified in both Males and Females.* *** Identified in both Females and subjects with airway obstruction.*
N = number of genes in the GO term. *DE = number of genes that are differentially methylated.*
P.DE = p-value for over-representation of the GO term. *FDR = False discovery rate.*

DNA methylation age

We calculated the biological age (DNAmAge) of the 1,561 subjects included in our cohort according to methods described by Horvath et al.[11] The biological age, calculated DNAmAge, was highly significantly associated with chronological age (r^2=0.806, p<0.001). Next, we calculated the age-acceleration by subtracting the chronological age from the DNAmAge. We tested if age-acceleration was associated with exposure to pesticides by performing a linear regression analysis, thereby adjusting for the potential confounders sex, age, smoking and pack years (Figure E3). Exposure to pesticides was included by two dummy variables for respectively low and high exposure to pesticides. We did not find any significant association between age-acceleration and exposure to pesticides (p_low = 0.196. p_high = 0.951).

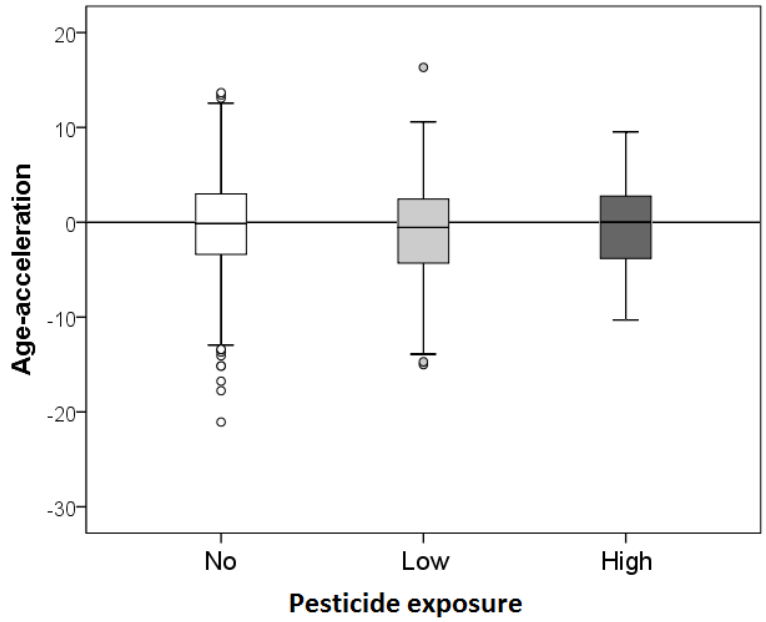

Figure E3: Boxplot of age-acceleration per pesticide exposure group.

References

1 Scholtens S, Smidt N, Swertz MA, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int J Epidemiol 2015;44:1172-80 doi:10.1093/ije/dyu229 [doi].

2 Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;30:1363-9 doi:10.1093/bioinformatics/btu049 [doi].

3 Chen YA, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 2013;8:203-9 doi:10.4161/epi.23470 [doi].

4 Pidsley R, Y Wong CC, Volta M, et al. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics 2013;14:293,2164-14-293 doi:10.1186/1471-2164-14-293 [doi].

5 Lehne B, Drong AW, Loh M, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol 2015;16:37,015-0600-x doi:10.1186/s13059-015-0600-x [doi].

6 Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 2012;13:86. doi:10.1186/1471-2105-13-86 [doi]

7 Aryee MJ, Jaffe AE, Corrada-Bravo H, et al.Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;30:1363–9. doi:10.1093/bioinformatics/btu049 [doi]

8 Forget P, Khalifa C, Defour J-P, et al.What is the normal value of the neutrophil-to-lymphocyte ratio? BMC Res Notes 2017;10:12. doi:10.1186/s13104-016-2335-5

9 Winans B, Humble MC, Lawrence BP. Environmental toxicants and the developing immune system: A missing link in the global battle against infectious disease? Reprod. Toxicol. 2011;31:327–36. doi:10.1016/j.reprotox.2010.09.004

10 Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics 2016;32:286–8. doi:10.1093/bioinformatics/btv560 [doi]

11 Horvath S. DNA methylation age of human tissues and cell types. Genome Biol 2013;14:R115. doi:10.1186/gb-2013-14-10-r115