Associations between persistent organic pollutants and endometriosis: A multipollutant assessment using machine learning algorithms

Komodo Matta, Evelyne Vigneau, Véronique Cariou, Delphine Mouret, Stéphane Ploteau, Bruno Le Bizec, Jean-Philippe Antignac, Germán Cano-Sancho

To cite this version:

Komodo Matta, Evelyne Vigneau, Véronique Cariou, Delphine Mouret, Stéphane Ploteau, et al.. Associations between persistent organic pollutants and endometriosis: A multipollutant assessment using machine learning algorithms. Environmental Pollution, 2020, 260, pp.114066. 10.1016/j.envpol.2020.114066 . hal-03191074

HAL Id: hal-03191074
https://hal.inrae.fr/hal-03191074

Submitted on 21 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0269749119355356
Manuscript_8b84c0e16f933d34555ce53ff83a2e36

Title: Associations between persistent organic pollutants and endometriosis: a multipollutant assessment using machine learning algorithms

Authors: Matta, Komodo[a]; Vigneau, Evelyne[b]; Cariou, Véronique[b]; Mouret, Delphine [a]; Ploteau, Stéphane [c]; Le Bizec, Bruno [a]; Antignac, Jean-Philippe[a]; Cano-Sancho, German[a]*

Affiliations:

[a] LABERCA, Oniris, INRAE, 44307, Nantes, France

[b] StatSC, ONIRIS, INRAE, Nantes, France

[c] Service de gynécologie-obstétrique, CIC FEA, Hôpital Mère Enfant, CHU Hôtel Dieu, Nantes, France

* Laboratoire d'Étude des Résidus et Contaminants dans les Aliments (LABERCA), Route de Gachet - La Chantrerie – BP 50707, 44307 Nantes Cedex 3, France.

Email : laberca@oniris-nantes.fr (Cano-Sancho, G)

Keywords:

Endometriosis, endocrine disrupting chemicals, persistent organic pollutants, machine learning, multipollutant modelling

© 2020 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

Abstract

Endometriosis is a gynaecological disease characterised by the presence of endometriotic tissue outside of the uterus impacting a significant fraction of women of childbearing age. Evidence from epidemiological studies suggests a relationship between risk of endometriosis and exposure to some organochlorine persistent organic pollutants (POPs). However, these chemicals are numerous and occur in complex and highly correlated mixtures, and to date, most studies have not accounted for this simultaneous exposure. Linear and logistic regression models are constrained to adjusting for multiple exposures when variables are highly intercorrelated, resulting in instable coefficients and arbitrary findings. Advanced machine learning models, of emerging use in epidemiology, today appear as a promising option to address these limitations. In this study, different machine learning techniques were compared on a dataset from a case-control study conducted in France to explore associations between mixtures of POPs and deep endometriosis. The battery of models encompassed regularised logistic regression, artificial neural network, support vector machine, adaptive boosting, and partial least-squares discriminant analysis with some additional sparsity constraints. These techniques were applied to identify the biomarkers of internal exposure in adipose tissue most associated with endometriosis and to compare model classification performance. The five tested models revealed a consistent selection of most associated POPs with deep endometriosis, including octachlorodibenzofuran, cis-heptachlor epoxide, polychlorinated biphenyl 77 or trans-nonachlor, among others. The high classification performance of all five models confirmed that machine learning may be a promising complementary approach in modelling highly correlated exposure biomarkers and their associations with health outcomes. Regularised logistic regression provided a good compromise between the interpretability of traditional statistical approaches and the classification capacity of machine learning approaches. Applying a battery of complementary algorithms may be a strategic approach to decipher complex exposome-health associations when the underlying structure is unknown.

Main findings capsule

Elastic-net provided a good compromise between the interpretability and performance, but applying a battery of complementary models may be best to support complex links between exposure and disease.

Introduction

Endometriosis is a hormone-dependent gynaecological disease characterised by the presence of endometrial tissue outside the uterine cavity and contributes to a number of non-specific symptoms, such as chronic pelvic pain, dysmenorrhea, dyschesia, dyspareunia, and often infertility (Eskenazi et al., 2002; Giudice, 2010; Sampson, 1927). The precise aetiology of endometriosis remains unclear but is likely multicausal, influenced by hormonal, genetic, and environmental factors. Evidence from epidemiological studies suggests a relationship between risk of endometriosis and exposure to some organochlorine persistent organic pollutants (POPs) like dioxin 2,3,7,8-Tetrachlorodibenzodioxin (TCDD), polychlorobiphenyls (PCBs) and organochlorine pesticides (OCPs) (Cano-Sancho et al., 2019), mechanistically supported by experimental evidence (Bruner-Tran et al., 2010; Bruner-Tran and Osteen, 2010; Matta et al., 2019). In a previous case-control study conducted in France (Ploteau et al., 2017), we found statistically significant associations between presence of deep endometriosis and concentrations of certain POPs in adipose tissue (AT), including 1,2,3,7,8-pentachlorodibenzodioxin (PeCDD), octaochlorodibenzofuran (OCDF), polybromodiphenylether (PBDE) 183, polybromobiphenyl (PBB) 153 or cis-heptachlor epoxide, among others. The approach previously used considered one pollutant at a time, with multivariable logistic regression adjusting for known and suspected confounding variables. This approach may prone to bias if associations are due to correlated coexposures. For this reason, multipollutant models are today encouraged to evaluate coexposure-outcome associations under collinear frameworks (Lenters et al., 2018; Weisskopf et al., 2018).

Collinearity is an acknowledged problem in analyses based on ordinary least squares (i.e. linear regression). It occurs when two or more predictor variables are highly correlated, as is often the case in datasets with mixtures of environmental chemical exposures. This may exacerbate variances due to model misspecification, especially when prior biological knowledge of underlying associations is not available (Schisterman et al., 2017). In the last decades, novel statistical methods and computational frameworks have emerged motivated by the challenges posed by air pollution mixtures, expanding the spectrum of available approaches to address the various data constraints (Bellinger et al., 2017; Stafoggia et al., 2017; Taylor et al., 2016). Overall, the models may be grouped by their capacities to reduce data dimensionality, select variables (identify risk variables within highly redundant and correlated variables) and group or cluster observations (Billionnet et al., 2012; Stafoggia et al., 2017). Among epidemiological studies, however, application of multipollutant approaches using biomarkers of exposure has been growing at a more modest pace. Some recent simulation studies have compared the performance of several multipollutant models to identify exposome-health associations with both continuous and dichotomous outcomes together with their interactions (Agier et al., 2016; Barrera-Gomez et al., 2017; Lenters et al., 2018; Sun et al., 2013). Results suggest there is no one-size-fits-all model and that model selection must be made on the basis of the data structures.

At the same time, novel high-throughput approaches in mass spectrometry and generation of large spectral datasets have also favoured the implementation of data mining pipelines and machine learning (ML) techniques in some exposome-health studies (Bellinger et al., 2017; Manrai et al., 2017). Despite this, many powerful ML methods like neural networks, support vector machines, and boosting algorithms remain underexplored but show promise in their computational capacity for classification and variable selection with highly complex data (Stingone et al., 2017; Zhao et al., 2019). These algorithms have the potential to assess individual variable associations while simultaneously adjusting for coexposures, addressing the issue of collinearity. Although these ML methods have seldom been applied in the context of environmental epidemiology, their emerging use in medical research and other epidemiological fields (i.e. genetics) suggest that their application may hold promise for the novel development of multipollutant exposure models (Bellinger et al., 2017; Deist et al., 2018; Roffman et al., 2018; Tomiazzi et al., 2019).

In this context, the objective of the present study was to apply and evaluate the performance of several ML methods in identifying the health status of patients. Following previous settings for systematic comparison of approaches in exposome-health research (Lenters et al., 2018), classification performance criterion is used for parameter tuning and model comparison.

Predictive capacity of models as an endpoint, however, has minor interest in this etiological research context. Instead, variable selection is sought as an endpoint, as a first step towards exploring complex biomarker-health associations within multidimensional and highly collinear frameworks. Exploratory data analysis using these models is thus conceived for a better understanding of the underlying structure of the biomarkers of exposure and the associations with endometriosis.

Methods and Materials

Study Population

This study draws upon a case-control study conducted in Pays-de-la-Loire, France between 2013 and 2015, focusing on a group of 80 persistent pollutants analysed in the AT of a sample population with and without endometriosis. Study design, recruitment, and methods have been previously reported (Ploteau et al., 2017). Briefly, the study enrolled a total of 99 women ages 18-45. Cases (n= 55) included women diagnosed with deep endometriosis (with surgical confirmation) and controls (n = 44) comprised a similar group of women present at the clinic for other gynaecological issues unrelated to endometriosis, surgically confirmed to not have endometriosis and displaying no related clinical symptoms (i.e. chronic pelvic pain, dysmenorrhea, dyspareunia, infertility). From both groups, cases and controls, 2 g of parietal AT (subcutaneous fat) samples were collected and stored at -80°C. Data were gathered pertaining to the diagnosis, anthropometric variables, and other potentially relevant factors such as age, body mass index (BMI), breastfeeding and parity. All participants signed an informed consent form approved by the Bioethics Committee of GNEDS (Groupe Nantais d'Éthique dans le Domaine de la Santé).

Exposure Assessment

Biomarkers of exposure were determined in adipose tissue, which is the most stable matrix for POP measurements reflecting long-term exposure (Cano-Sancho et al., 2019). These exposure estimates capture the window between onset of the first symptoms and the diagnosis of endometriosis (7-10 years). The supporting methods used for chemical analyses have been published elsewhere (Antignac et al., 2009; Bichon et al., 2015; Ploteau et al., 2016; Ploteau et al., 2017). In brief, samples were quantified with 13C-labeled congeners using isotope dilution and extracted under high temperature and pressure (ASE Dionex, Sunnyvale, CA, USA). Gravimetric methods were used to measure fat content, and extracts were reconstituted in hexane for cleanup. OCPs were isolated using gel permeation chromatography; other target substances were isolated using three successive purification steps: acid silica, Florisil®, and celite/carbon columns. PCDD/F, PCB, PBDE, PBB and OCP were measured by gas chromatography (Agilent 7890A) coupled with high-resolution mass spectrometry (GC-HRMS) on double sector instruments (JEOL MS 700D and 800D) after electron impact ionization (70 eV), operating at 10000 resolutions (10% valley) and in the single ion monitoring (SIM) acquisition mode. HBCD isomers were quantified using liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) on a triple quadrupole instrument (Agilent 6410) using electrospray ionization and selective reaction monitoring. The full list of analysed chemicals and congeners can be ground in the Supplemental Table S1. All methods were validated according to Regulation (EU) No 376/2014 of the European Parliament (EU, 2014). Analysis was performed in an ISO 17025:2005 accredited laboratory. All internal exposure data were generated blinded to the case/control status of samples. Recoveries were in the 80–120% range, and expanded uncertainty was lower than 20%. Exposure levels for POPs were expressed in a lipid-weight basis (lw).

Data pre-processing

Missing data were characterised to determine their nature (Missing at Random vs Missing Not at Random). Numerical covariates missing at random (i.e. BMI, age) were imputed using MICE package in R. Distributions before and after imputation were checked to ensure consistency. Data missing not at random included several POPs that were either not detected through quantification or were found to be below the limit of detection. Exposure variables lower than the limit of detection (LOD) were assigned a value of LOD/2 (Cohen and Ryan, 1989). Variables for which over 75% of exposure data were missing or below LOD were excluded from analysis for quality control purposes (See Table S1). Remaining exposure variables were log transformed, centred and scaled by their standard deviations.

Exploratory Data Analysis

Distributions of exposure levels of chemicals from cases and controls were summarised by median and interquartile ranges, and compared statistically by using Mann-Whitney-Wilcoxon tests. For all data analyses, the significance level threshold was set to $p < 0.05$.

A first multivariate exploratory analysis was performed to investigate and visualise the underlying structure of the exposure data matrix. Bivariate correlation analysis was performed using Spearman rank test and depicted in heatmaps. Principal Component Analysis (PCA), run with *FactoMineR* package in R, and Clustering of variables around Latent Variables (CLV), run with *ClustVarLV* package, were used to detect clusters of co-observed exposure variables (Vigneau et al., 2015). Similar to PCA, CLV latent variables associated with clusters are synthetic components to facilitate data variability description.

Multipollutant Data Analysis

For multipollutant analysis, five supervised algorithms for classification and variable selection were applied (Regularised logistic regression, Artificial Neural Network (ANN), Support Vector Machine (SVM), Adaboost (ADA), and Partial Least Squares Discriminant Analysis (PLSDA). All models were run in a full mode (wherein all variables are included into the model) and with sparsity constraints (wherein classification performance is used to select only the most discriminant variables to include in the model). Sparse models, shrink the weight of less discriminant variables to zero, thus simplifying the model for classification purposes and addressing the risk of overfitting. For algorithms without inherent sparsity constraints, we employed Recursive Feature Elimination (RFE), a resampling approach that selects the subset of variables that minimises the model classification error by iteratively removing one feature at a time. Briefly, RFE follows three steps: (1) training the classifier by optimising feature weights; (2) computing the ranking criterion for all features, and finally (3) removing the feature with the smallest ranking criterion (Guyon, 2002; Kuhn, 2008). The process is then repeated.

Data was randomly partitioned in an 80/20 ratio for a training set and test set. The training set comprised 80% of observations and was used to train the algorithm to better understand the exposure profile of individuals with and without endometriosis (endometriosis status known). The test set, which comprised the remaining 20%, was used to evaluate the classification performance of the trained algorithm.

Parameters were optimised for efficiency using a ten-times repeated cross validation (CV) to exhaust the dataset. Tuning parameters were calibrated and set for each model individually.

For each model the coefficients associated with all (full models) or selected variables (sparse models) were estimate, generating a weight, or importance, according to its contribution to the final model. Thus, variables with greater variable importance values (VI) corresponded to those which contribute more to the final model.

We also computed metrics for classification performance including Receiver Operating Characteristic (ROC), Area Under the Curve (AUC), sensitivity, and specificity. ROC curves measure a test's ability to discriminate between cases and controls and is quantified by the AUC. An AUC of 1 means the test has 100% discriminative capacity, and a value of 0.5 means the test is unable to discern cases from controls any more than random chance. In general, values between 0.9-1.0 are considered very good, values 0.8-0.9 are considered good, 0.7-0.8 as fair, 0.6-0.7 as poor, and 0.5-0.6 as failure (Tape, 2001). Sensitivity measures the capacity of the model to correctly identify positive cases, while specificity indicates the capacity to correctly identify controls. McNemar's test on paired proportions was used to assess the predictive accuracy of the classification model. We also compared the agreement of variables selected between models, their VI, their interpretability and flexibility to be applied in epidemiological studies.

All statistical analyses were performed in R software v.3.4.3. Model performance evaluation was conducted with the R Caret framework (Kuhn, 2008) that links multiple packages and functions for modeling, specifications summarized in Table 2.

a) Regularised logistic regression: ridge and elastic-net regression

Elastic-net (ENET) is a penalised regression model, which integrates generalised regression models with regularisation techniques using penalty functions. It combines ridge regression, which applies a penalty term to the sum of squared coefficients to favour grouping highly correlated predictors, and a lasso constraint on the sum of the absolute values of the coefficients to minimise the impact of irrelevant variables and set their coefficients to zero. This provides the model sparsity (lasso) and robustness (ridge) (Zou and Hastie, 2005). The final model will thus include fewer features than the initial state, which is helpful to avoid overfitting the model to the training data. For this reason, ENET is particularly adapted to variable selection of data with high collinearity (Lenters et al., 2016).

ENET is implemented using the *glmnet* function of the R package *glmnet*. Tuning parameters of *glmnet* are *alpha* (lasso, mixing percentage) and *lambda* (regularisation parameter). Alpha and lambda values ranged from 0 to 1. For the full model, the lasso penalty term *alpha* was set to 0, thus eliminating its intrinsic sparsity parameter.

b) Artificial Neural Network

ANNs are ML algorithms inspired by the structure of biological neural networks. They consist of a number of interconnected neural nodes. The structure of ANNs usually comprise three principal layers: the input layer includes input nodes (predictor variables), the output layer consists of a single output node (endometriosis status), and the middle hidden layer(s) are populated by a collection of hidden nodes with values which model the complex relationships between the input and output layers but which do not of themselves have a real world analogue. The synapses which connect each of these layers' nodes to one another are weighted, which represents the strength of the connection, similar to coefficients in logistic regression models. In neural networks, the weight decay value acts as the regularisation term.

ANN is implemented using the *nnet* package. Tuning parameters are *size* (number of hidden layers) and *decay* (weight decay). Size ranged from 1 to 50, and decay from 0 to 0.9.

c) Support Vector Machine

SVM is a classifier that works by reimagining data in a multidimensional space and generating multiple potential hyperplanes to separate data, then selecting the optimal hyperplane which maximises the margins between the two groups (here, cases and controls). Typically, SVMs are used as a linear classification model, but they can also generate hyperplanes for nonlinear data using a kernel function. In this study, we used an SVM with a radial basis function (RBF) kernel for nonlinear data to transform the original feature space for better separation of the two groups. Regularisation is controlled by a cost parameter. The cost parameter controls the tradeoff between training errors and model complexity. A smaller cost value increases the number of training errors while larger costs may lead to overfitting. The sigma parameter with RBF kernel determines the flexibility of the decision boundary and how much influence a single feature can exert. Larger sigmas create a more flexible and smooth decision boundary with more variance and thus act as a more general classifier, while smaller sigma values are stricter and tend to make more local classifiers (Ben-Hur A., 2010). Tuning parameters of *svmradial* from the *kernlab* package are *sigma* (Sigma) and *C* (Cost). Sigma ranged from 0.001 to 1, and cost ranged from 0 to 100.

d) Boosting trees: Adaboost

Boosting algorithms iteratively combine the output of multiple weaker classifiers (decision trees) in a stepwise manner to improve performance at each iteration to make a strong classifier. Combining the boosting technique with decision trees allows each subsequent iteration to focus on increasingly harder to classify observations, regularising iteratively, and ultimately yielding a weighted sum which serves as the final classifier. Individual decision trees that are more performant contribute more to the final classifier. In this study, we used adaptive boosting, Adaboost (ADA), which specialises in minimising exponential loss function by adapting the weights to increase accuracy in predictions (Friedman et al., 2000). ADA Classification Trees was computed with the package *fastAdaboost* with tuning parameters *nIter* (number of trees), which ranged from 10 to 500, and *method* (boosting method).

e) Partial least squares discriminant analysis

Partial least squares discriminant analysis (PLSDA) models approximate the relationship between predictor variables and the response variable (endometriosis status), searching for directions of maximum covariance between the two. Using the *softmax* function, predictor variables are assigned "probability-like" values (on a scale of 0 to 1 which sum to 1), and the class with the largest class probability is the predicted class. In the sparse form, only the most predictive or discriminative features from the data are selected to inform classification.

The tuning parameter of *plsda* from the *pls* package is *ncomp* (number of components), which ranged from 2 to 54.

Results

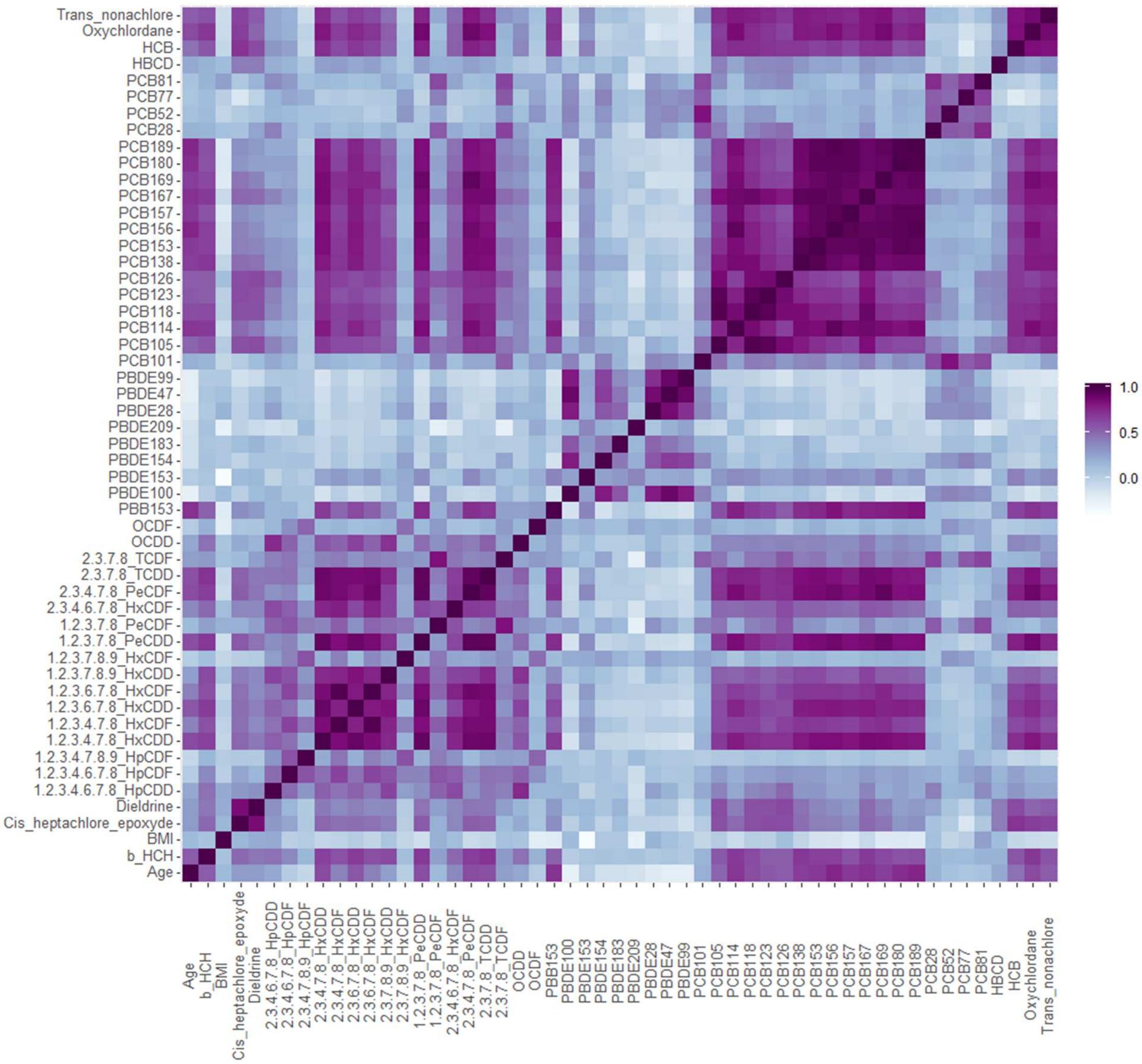
Descriptive analysis

Cases (n = 55) and controls (n = 44) were matched for age, BMI, and breastfeeding history, three factors which are known to be strongly correlated with internal exposure levels of POPs (Ploteau et al., 2016). Mean and standard deviation age of control and case group were 32.6 (±6.5) and 34.3 (±6.2) years, respectively (Student T test, p=0.19). BMI also did not differ between groups, with 25.4 (±5.9) kg/m^2 for controls and 24.0 (±5.1) kg/m^2 for cases (p=0.21). Parity and breastfeeding were not included in the models due to their uncertain causal role in the pathogenesis of endometriosis (Ploteau et al., 2017; Upson et al., 2013). Cases exhibited lower average breastfeeding duration (4.1±14.9 months) than controls (1.3±3.1 months), but did not differ statistically (p=0.18). Distributions of concentrations of POPs in AT for cases and controls are provided in Supplemental Table S2.

Exploratory Data Analysis

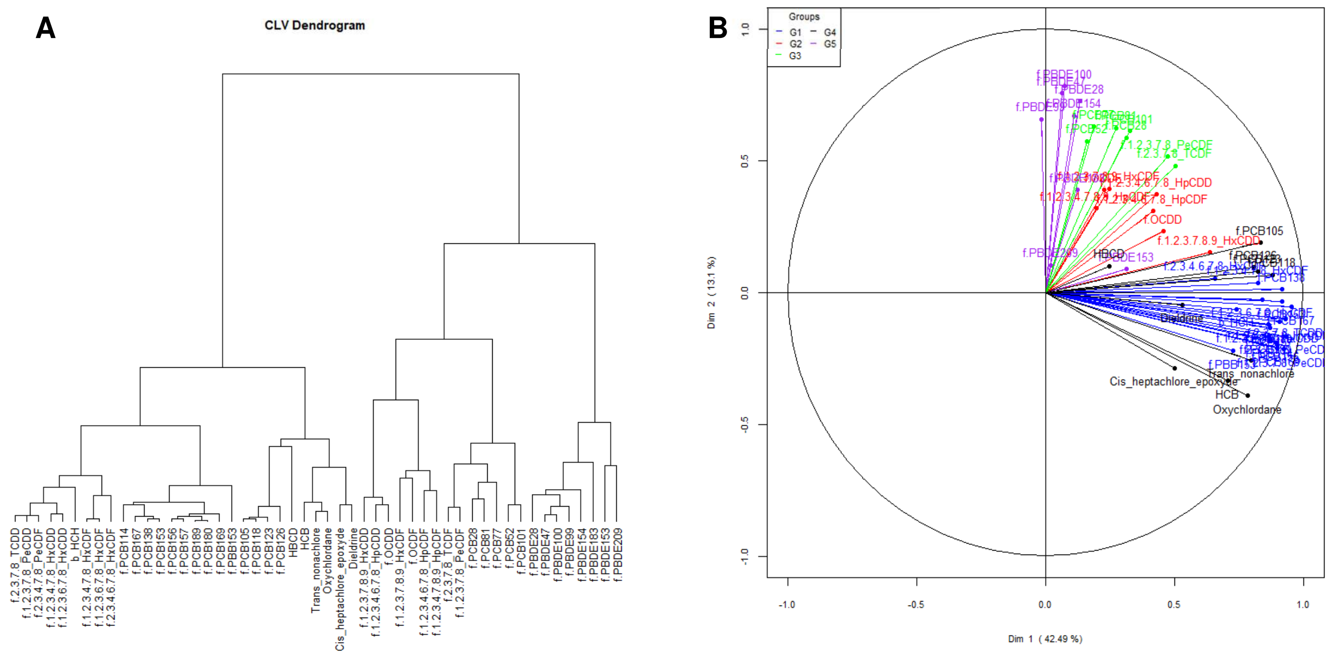
Coefficients from the bivariate correlation analysis between pollutants are depicted in the heatmap in Figure 1. Clusters of dioxins, PCBs, brominated flame retardants and pesticides present positive correlations. Coplanar PCBs 189, 169, 167, 157, 156, 126, 123, 118, 114, 105 were found to be positively correlated amongst one another but not with coplanar PCBs 77 and 81. Interestingly, OCDF was not found to be strongly correlated with any other variable, save for a moderate positive association with 1.2.3.7.8.9 HxCDF and 1.2.3.4.7.8.9 HpCDF. PBDEs were found to be mildly negatively correlated with dioxins, furans, and pesticides, 1.2.3.7.8 PeCDF and 2.3.7.8 TCDF, and non-coplanar PCBs 28, 52, and 101. Age was mildly positively correlated with the same clusters of dioxins and coplanar PCBs. BMI did not show any correlations with any of the other variables. Heatmaps displaying the correlation analysis stratified by endometriosis status did not show visual differences between cases and controls (Supplemental Figure S1).

Figure 1. Correlation analysis heatmap



With regard to PCA, the two first components summarise more than a half of the data (42.49% of inertia retrieved by the first component, 13.10% by the second) (Supplemental Figure S2A). Factor maps depicting the correlations between pollutants variables and the two components are available in Supplemental Figures S2B-C.

Figure 2. Clustering of the exposure variables using CLV, (A) Dendrogram and (B) representation of the partition into five clusters on the basis of the two dimensional PCA variables configuration.



CLV revealed the underlying structure of the data, which can be visualised in a dendrogram (Figure 2A), five clusters (K = 5) of which can be seen in a two dimensional PCA variables configuration (Figure 2B). The groups identified tend to form clusters around extant chemical families: dioxins, furans, pesticides, coplanar PCBs, non-coplanar PCBs, PBDEs, and PBBs. The partition of variables has been defined so that within each cluster the angles between vectors associated with the exposure variables and a latent (not observed) central variable are minimised (maximising correlation). However, some exposure variables such as HCBD (G4), or PBDE209 and PBDE153 (G5) which are both far from the centre of their respective cluster and not well represented into the first PCA map may have been difficult to assign to any of the five clusters highlighted. In the dendrogram, it can be seen that HCBD would be in its own cluster at K = 11, and that PBDE209 and PBDE153 form a very small cluster.

Multipollutant Data Analysis

Parameter optimisation plots are available in Supplemental Figures S3-S7 and the final selected parameters are summarised in Table 1.

Table 1. Summary of algorithms, package functions and parameters optimised throughout the calibration process for the full and sparse models.

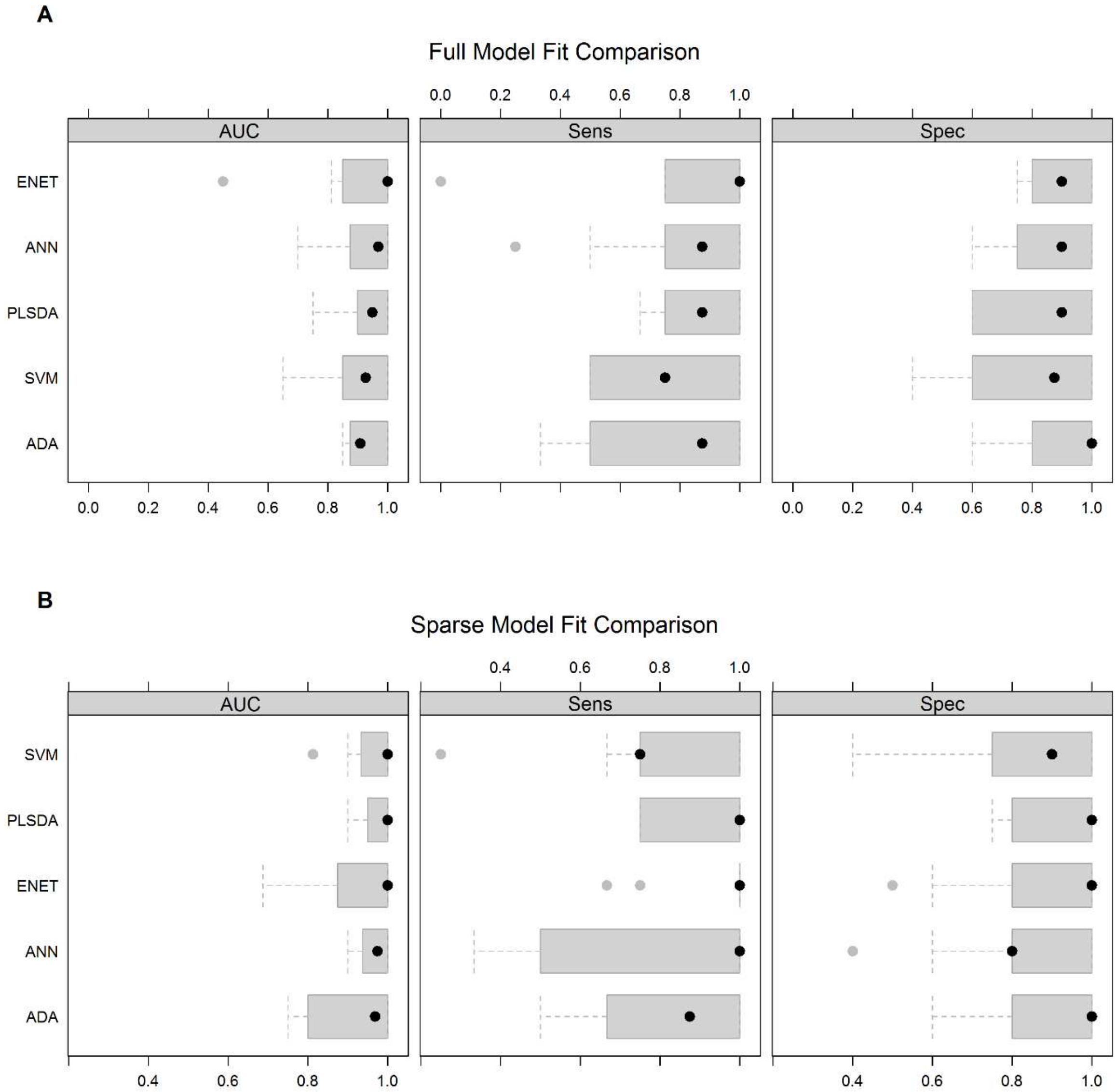
Model	Package	Method	Tuning Parameters (full)	Tuning Parameters (sparse)
Regularised logistic regression	*glmnet*	*glmnet*	alpha = 0 lambda = 0.05	alpha = 0.3 lambda = 0.1
Artificial Neural Network	*nnet*	*nnet*	size = 2 decay = 0.8	size = 2 decay = 0.8
Support Vector Machine	*kernlab*	*svmRadial*	sigma = 0.001 C =100	sigma = 0.001 C = 100
Adaboost	*fastAdaboost*	*adaboost*	nIter = 100 method = Adaboost.M1	nInter = 100 method = Adaboost.M1
Partial Least Squares - Discriminant Analysis	*pls*	*plsda*	ncomp = 5	ncomp = 2

Full Models

For each model, a list of VIs was generated, signifying to what extent each variable contributed to the final model (Figure S8).

Models were further compared according to their fit (Figure 3A, Table S3) and classification performance (Table S4) using a confusion matrix to determine accuracy, AUC, sensitivity, and specificity. Ridge, SVM and ANN scored highest in AUC (SD) (0.968 (0.035), 0.958 (0.059), 0.956 (0.063) respectively. ENET had the highest scoring sensitivity (0.900 (0.129)) and ANN had the highest scoring specificity (0.900 (0.175)) with the lowest sensitivity (0.775 (0.208)).

Figure 3. Model Fit Comparison for (A) Full and (B) Sparse Models presented in median and interquartile range.



Sparse Models

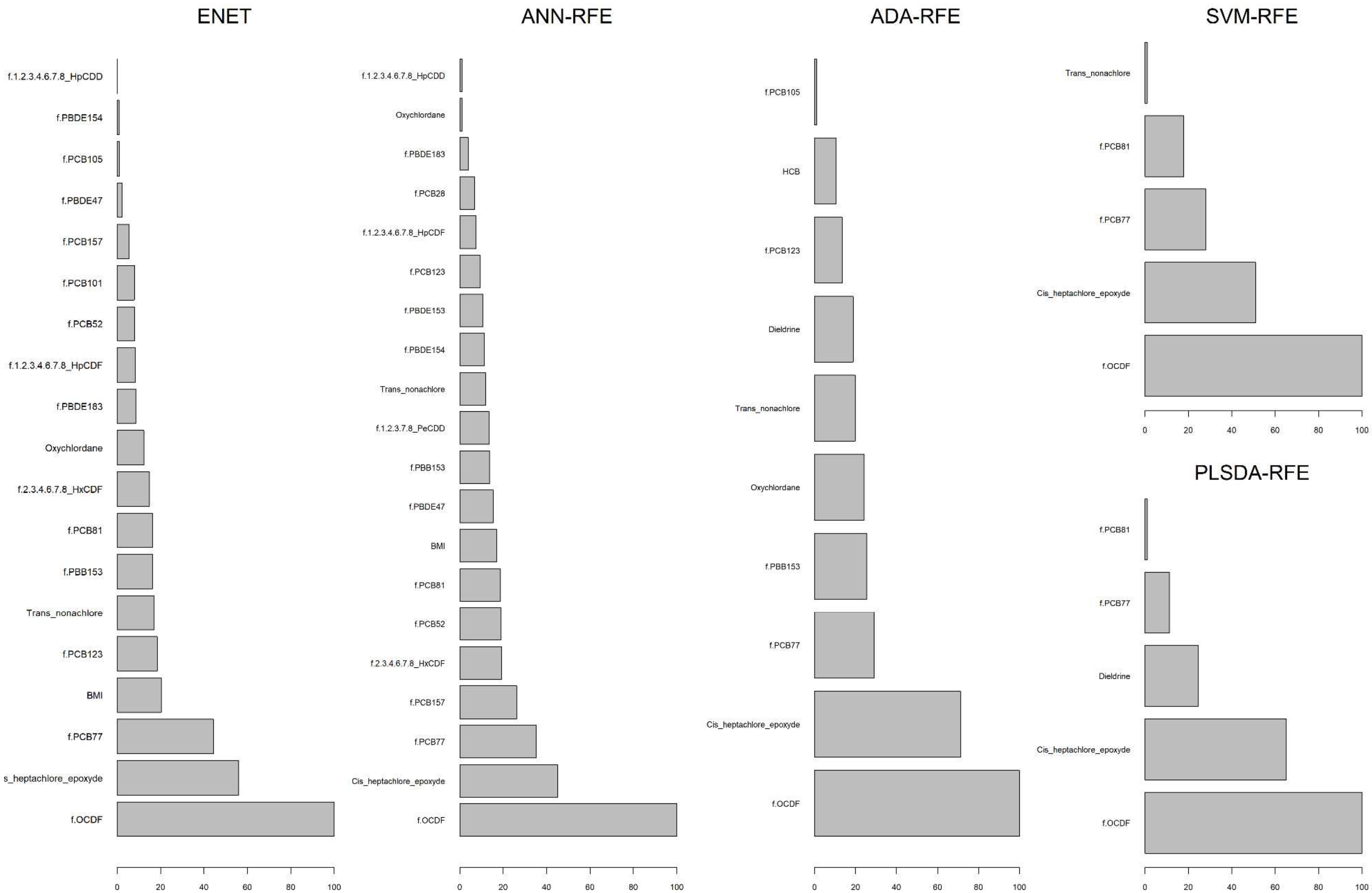
Calibration plots of variable selection for each model are available in Supplemental Figures S9-S13. Nineteen variables were identified by ENET (OCDF, cis-heptachlor epoxide, PCB77, PCB81, BMI, PCB123, trans-nonachlor, PCB52, PCB101, PCB157, 2.3.4.6.7.8 HxCDF, PBB153, 1.2.3.4.6.7.8 HpCDF, Oxychlordane, PBDE183, PBDE154, and 1.2.3.4.6.7.8 HpCDD); twenty variables were identified by ANN (OCDF, cis-heptachlor epoxide, PCB77, PCB81, PBB153, BMI, 2.3.4.6.7.8 HxCDF, PCB157, 1.2.3.7.8 PeCDD, PBDE154, PBDE47, PCB52, trans-nonachlor, PCB28, PBDE153, PCB123, 1.2.3.4.6.7.8 HpCDD, oxychlordane, PBDE183, 1.2.3.4.6.7.8 HpCDF); five were identified by SVM (OCDF, cis-heptachlor epoxide, PCB77, PCB81, trans-nonachlor); ten were identified by ADA (OCDF, cis-heptachlore epoxide, PCB77, PBB153, oxychlordane, trans-nonachlor, dieldrin, PCB123, HCB, PCB105) and five by PLSDA (OCDF, cis-heptachlor epoxide, dieldrin, PCB77, and PCB81) (Figure 4).

Of particular interest, three variables were identified by all five models (OCDF, cis-heptachlor epoxide, and PCB77). Trans-nonachlor and PCB81 were identified by four of the five models. Three of the models identified PBB153, PCB123, and oxychlordane as important variables.

Summary of classification performance metrics (accuracy, AUC, sensitivity, and specificity) are presented in Figure 4B and Table S4 for each model. Model fit accuracy for sparse models all ranged from 85.0-88.8%, and AUC indices (SD) were all greater than 0.95 (ENET 0.988 (0.024), ANN 0.989 (0.024), SVM 0.973 (0.058), ADA 0.954 (0.039), PLSDA 0.980 (0.045)). Sensitivity across models did not vary markedly from one another (ENET 0.817 (0.211), ANN 0.900 (0.129), SVM 0.891 (0.142), ADA, 0.867 (0.188), PLSDA 0.975 (0.079)), nor did specificity (ENET 0.915 (0.111), ANN 0.895 (0.146), SVM 0.885 (0.256), ADA, 0.870 (0.106), PLSDA 0.775 (0.203)). Values of all model fit metrics are available in Supplemental Table S3.

Finally, statistical significance of paired proportions was calculated in a confusion matrix. ENET, ANN, SVM, and ADA with RFE had a prediction accuracy of 84.2% ($p = 0.015$), which was significantly better than chance (57.9%); on the contrary, PLSDA with feature selection failed in significantly classifying better than chance (Figure S14). Sensitivity and specificity are listed in Supplemental Table S4.

Figure 4. Variables selected for sparse models on a 0-100 scale of predictive relative importance



Discussion

In this study, we applied for the first time a selection of multipollutant models, including three ML classifiers scarcely used in epidemiology, to support variable selection from a highly correlated dataset of POPs biomarkers. Full and sparse models were investigated to compare the balance between bias, variance, classification performance and interpretability of results. Full models, which include every variable into the final model, may be more useful in terms of biological interpretation, but at the risk of being computationally cumbersome, overfitting the data, and including unnecessary variables, especially when dealing with high dimensional data. Sparse models, which select variables on the basis of minimising classification error, address the issues of dealing with high dimensional data, but may fail to reveal true underlying biological associations by selecting only one representative biomarker from a cluster of correlated variables, as one particularly strong association may mask other structurally associated predictors. It is thus important in sparse models to note not only which variables are commonly selected across models but also which differ, taking into account their bivariate relationships as well. Thus in order to support the biological interpretation of findings and taking advantage of both types of models, the variables identified from sparse models should be judged against the structures from full models and the interdependency between variables. In any case, variable selection should be considered as a preliminary step to support the construction of causal structures and explanatory models under high dimensional settings with correlated exposures, as commonly found with POP-endometriosis research.

This initial exploration supports the use of regularised regression (i.e. elastic-net) for variable selection, exhibiting an adequate balance between classification performance and interpretability. In this study, powerful classifiers such as SVM, ANN or ADA did not outperform other commonly used algorithms such as PLSDA or ENET. Globally, variable selection was very consistent across the different models with minor differences in the biomarker rankings. For sparse models, the number of discriminant variables retained was substantially lower for SVM and PLSDA than for the other models. Three variables appeared as the strongest predictors of endometriosis status, namely OCDF, cis-heptachlor epoxide, and PCB77. Trans-nonachlor and PCB81 were identified by four of the five tested models, while PBB153, PCB123, and oxychlordane were identified by three. These results are consistent with our previous findings using a sequential logistic regression followed by false discovery rate correction, with an Odds Ratio (95% CI) of 5.42 (2.73-12.85) and 5.36 (2.44-14.84) for OCDF and cis-heptachlor epoxide, respectively (Ploteau et al., 2017). Coplanar PCB 77 and 123, as well as polybrominated flame retardant PBB153, were also identified as important predictors. The correlations between pesticides cis-heptachlor epoxide and trans-nonachlor with PCDDs, coplanar PCBs, several furans and non-coplanar dioxins might mask the impact of the latter on endometriosis in sparse models. Interestingly, OCDF, the strongest signal identified by all five models, was not strongly correlated with any other predictor variable.

The model fit of five models did not differ substantially in terms of AUC, specificity, or sensitivity. In this study, all five models had AUC values greater than 0.9, suggesting that this battery of algorithms presents a promising method of modelling the associations between concentrations of POPs in AT and endometriosis status. Interestingly, PLSDA with RFE performed well in model fit (AUC = 0.98) but scored lowest in classification accuracy (i.e. 0.68 (95% CI; 0.43, 0.87)). This may be due to the use of RFE to induce sparsity, instead of using the intrinsic sparse PLSDA (sPLDSA) with lasso penalisation of PLS loading vectors (Le Cao et al., 2011; Le Cao et al., 2008). The performance of the multiclass wrapper RFE has shown to decrease dramatically with the number and correlation of variables due to the backward elimination used for variable selection (Le Cao et al., 2011). We have applied RFE here to allow direct comparison among models; however, future studies with wide and highly correlated datasets should consider the use of sPLSDA over the RFE procedure. Surprisingly, powerful classifiers such as ANN and SVM did not outperform the classification performance of more standard methods such as ENET with the present dataset. The small sample size of the dataset might explain the imperfect architecture of hidden layers, the number of neurons in each layer, and the activation functions in ANN (Alwosheel et al., 2018).

Despite the emergent use of multipollutant models in environmental epidemiology, few studies have applied ML algorithms to gain better insight into the complex exposome-health associations (Stafoggia et al., 2017). In the field of endometriosis, two previous multipollutant approaches have addressed high-dimensional POP biomarker data structures from a common case-control study (Louis et al., 2005). The first study (Roy et al., 2012) applied a data-driven reduction approach, Bayesian Belief Network, to identify the most associated biomarkers conditional to all other exposures and including biologically relevant covariates of endometriosis. Authors found PCB114 as the most influential biomarker from a mixture of 62 congeners. The second (Zhang et al., 2012) applied latent class models for a joint analysis of PCB mixtures, characterising biomarker-specific differences through random effects, accommodating the number of ordinal latent classes. Additionally, several recent studies have employed batteries of ML models to study other risk factors on health outcomes. For instance, Zhao et al. (2019) tested four different algorithms (ANN, SVM, ADA, Random Forest (RF)) on a population of 1113 workers exposed to industrial noise to predict hearing impairment. Predictive accuracy was found to be between 78.6-80.1% for all four models, which is comparable to the accuracies for ANN, SVM, and ADA (84.2%) found in this study. Although SVM had slightly higher accuracy than the other three models, the predictive abilities of the four models were not significantly different. Authors concluded that these algorithms may be a feasible tool for evaluation and prediction. Tomiazzi et al. (2019) evaluated hearing impairment in 127 Brazilian farmers exposed to pesticides and/or cigarette smoke, using ANN, SVM, and K-Nearest Neighbour. The models were able to distinguish exposure group from control group but failed to differentiate between five different exposure classes (Tomiazzi et al., 2019).

Nevertheless, some methodological limitations remain. One challenge of ML algorithms is the balance between model complexity and classification performance. Full models, which can be powerful tools in mapping relationships between predictors and outcome, may overfit the data, as every variable is included in the final model even if they are arbitrary noisy variables. Sparse models risk losing valuable biologically relevant information in favour of predictive performance. There is currently no consensus on how to measure degree of overfitting, despite the intensive use of validation techniques aimed at controlling such risk (Hastie, 2009). Model performance depends heavily on not only the size of the datasets but also on the parameters of each model. Simulation studies have shown little impact of sample size on classification performance of ENET, lasso, boosted trees or sPLSDA, in high-dimensional (p=50) and high-correlated datasets (ρ=0.8) (Lenters et al., 2018). Nonetheless, our findings should be carefully considered due to the small number of observations of the dataset (n = 99). Sample size may also impact the stability of coefficients and the reproducibility of results, an inherent issue of data-driven calibrations based on k-fold CV to select the tuning parameters (Lim and Yu, 2016). Furthermore, we only conducted internal CV for model optimisation and model performance evaluation, constraining the generalisability of our findings and highlighting the need for supplementary analogue datasets to externally validate the findings.

As the variable selection process should be considered a preliminary step previous to inferential analysis, an additional challenge posed by ML is the interpretability of outputs. ML algorithms are often viewed as “black boxes,” where it is difficult to inspect the inner workings of how outputs are generated and what they mean in a real-world context. The coupling of modelling techniques with graphical approaches has been proposed as a crucial way to apply and interpret ANNs in epidemiological research (Duh et al., 1998). In a simulation setting, kernel mapping in combination with a perceptron neural network has shown to efficiently generate odds ratios from perceptron weights to ease epidemiological interpretation of complex nonlinear exposure-disease associations (Heine et al., 2011). Future simulation studies should aim to extend the knowledge of model performance of ML classifiers in exposome-health settings, exploring the impact of parametrisation, sample size, correlation and interaction between exposure variables (Barrera-Gomez et al., 2017; Lenters et al., 2018).

The field of biomarkers for exposure assessment is moving fast towards a more chemical agnostic paradigm, favouring the generation of massive spectral datasets (Andra et al., 2017). Application of this novel high-throughput technology in epidemiology will demand an accommodation of epidemiological frameworks and clear harmonisation and standardisation of statistical workflows for comparability of findings (Manrai et al., 2017). Thus, novel approaches should empower multidimensional modelling to account for confounding and mediation of biomarker mixtures (Bellavia et al., 2019; Mostafavi et al., 2019). For instance, two-stage regression has been applied to address confounding, with a preliminary regression step between each outcome and exposure against the confounders, and a secondary sPLS regression fitting the resulting residuals (Lenters et al., 2015). The targeted maximum-likelihood based estimation is a doubly robust approach with powerful applications in causal inference of observational research. This approach has the potential to integrate multiple environmental and dietary exposures with confounding variables (Papadoupoulou et al., 2019). Considering that there is no one single algorithm with a definitive approach to build multipollutant models in exposome-health associations, the statistical exposome toolbox should be furnished with a variety of complementary algorithms to support the understanding of complex associations. In this regard, these novel ML algorithms seem a promising complement to characterising non-linear associations under highly collinear circumstances, especially in cases were the interpretability may be compromised in favour of identifying subtler statistical signals from noise (Hamra and Buckley, 2018).

Conclusions

In conclusion, the tested ML models were able to consistently reveal a number of pollutants associated with endometriosis, including OCDF, heptachlor epoxide and PCB77. The high classification performance for all five models suggests that ML may be a promising complementary approach in modelling highly correlated exposure matrices and their associations with health outcomes. It is important, however, to perform a follow-up explanatory statistical analysis on the identified variables of interest to make biological inferences. Regularised logistic regression provided a good compromise between the interpretability of traditional statistical approaches and the classification capacity of machine learning approaches for this initial exploration. Applying a battery of complementary algorithms may be a strategic approach to decipher complex exposome-health associations when the underlying structure is unknown. Future simulation studies should aim to evaluate the impact of parametrisation, overfitting, sample size, correlation between variables and to quantify model stabilities.

Declaration of Interest

Authors declare no conflicts of interest.

Acknowledgements

KM received a French regional government grant for doctoral allocations Pays de la Loire CPER 2014-2020.

References

Agier, L., Portengen, L., Chadeau-Hyam, M., Basagana, X., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., Gonzalez, J.R., Nieuwenhuijsen, M.J., Vineis, P., Vrijheid, M., Slama, R., Vermeulen, R., 2016. A Systematic Comparison of Linear Regression-Based

Statistical Methods to Assess Exposome-Health Associations. Environ Health Perspect 124, 1848-1856.

Alwosheel, A., van Cranenburgh, S., Chorus, C.G., 2018. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. Journal of Choice Modelling 28, 167-182.

Andra, S.S., Austin, C., Patel, D., Dolios, G., Awawda, M., Arora, M., 2017. Trends in the application of high-resolution mass spectrometry for human biomonitoring: An analytical primer to studying the environmental chemical space of the human exposome. Environ Int 100, 32-61.

Antignac, J.-P., Cariou, R., Zalko, D., Berrebi, A., Cravedi, J.-P., Maume, D., Marchand, P., Monteau, F., Riu, A., Andre, F., Le Bizec, B., 2009. Exposure assessment of French women and their newborn to brominated flame retardants: Determination of tri- to deca-polybromodiphenylethers (PBDE) in maternal adipose tissue, serum, breast milk and cord serum. Environmental Pollution 157, 164-173.

Barrera-Gomez, J., Agier, L., Portengen, L., Chadeau-Hyam, M., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., Gonzalez, J.R., Nieuwenhuijsen, M., Vineis, P., Vrijheid, M., Vermeulen, R., Slama, R., Basagana, X., 2017. A systematic comparison of statistical methods to detect interactions in exposome-health associations. Environ Health 16, 74.

Bellavia, A., James-Todd, T., Williams, P.L., 2019. Approaches for incorporating environmental mixtures as mediators in mediation analysis. Environ Int 123, 368-374.

Bellinger, C., Mohomed Jabbar, M.S., Zaiane, O., Osornio-Vargas, A., 2017. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health 17, 907.

Ben-Hur A., W.J., 2010. A User’s Guide to Support Vector Machines. Data Mining Techniques for the Life Sciences 609, 223-239.

Bichon, E., Guiffard, I., Venisseau, A., Marchand, P., Antignac, J.P., Le Bizec, B., 2015. Ultra-trace quantification method for chlordecone in human fluids and tissues. J Chromatogr A 1408, 169-177.

Billionnet, C., Sherrill, D., Annesi-Maesano, I., 2012. Estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol 22, 126-141.

Bruner-Tran, K.L., Ding, T., Osteen, K.G., 2010. Dioxin and endometrial progesterone resistance. Semin Reprod Med 28, 59-68.

Bruner-Tran, K.L., Osteen, K.G., 2010. Dioxin-like PCBs and endometriosis. Syst Biol Reprod Med 56, 132-146.

Cano-Sancho, G., Ploteau, S., Matta, K., Adoamnei, E., Louis, G.B., Mendiola, J., Darai, E., Squifflet, J., Le Bizec, B., Antignac, J.P., 2019. Human epidemiological evidence about the associations between exposure to organochlorine chemicals and endometriosis: Systematic review and meta-analysis. Environment International 123, 209-223.

Cano-Sancho, G., Marchand, P., Le Bizec, B., Antignac, J.P., 2020. The challenging use and interpretation of blood biomarkers of exposure related to lipophilic endocrine disrupting chemicals in environmental health studies. Molecular and Cellular Endocrinology 1;499:110606.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Janssen, N.A.H., Martin, R.V., Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B., Hoek, G., 2019a. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environment International 130, 104934.

Cohen, M.A., Ryan, P.B., 1989. Observations Less than the Analytical Limit of Detection: A New Approach. JAPCA 39, 328-329.

Duh, M.S., Walker, A.M., Ayanian, J.Z., 1998. Epidemiologic interpretation of artificial neural networks. Am J Epidemiol 147, 1112-1122.

Eskenazi, B., Mocarelli, P., Warner, M., Samuels, S., Vercellini, P., Olive, D., Needham, L.L., Patterson, D.G., Jr., Brambilla, P., Gavoni, N., Casalini, S., Panazza, S., Turner, W., Gerthoux, P.M., 2002. Serum dioxin concentrations and endometriosis: a cohort study in Seveso, Italy. Environ Health Perspect 110, 629-634.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann. Statist. 28, 337-407.

Giudice, L.C., 2010. Clinical practice. Endometriosis. N Engl J Med 362, 2389-2398.

Guyon, I., Weston, J., Barnhill, S. et al., 2002. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46, 389–422.

Hamra, G.B., Buckley, J.P., 2018. Environmental exposure mixtures: questions and methods to address them. Curr Epidemiol Rep 5, 160-165.

Hastie, T., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

Heine, J.J., Land, W.H., Egan, K.M., 2011. Statistical learning techniques applied to epidemiology: a simulated case-control comparison study with logistic regression. BMC Bioinformatics 12, 37.

Jain, P., Vineis, P., Liquet, B., Vlaanderen, J., Bodinier, B., van Veldhoven, K., Kogevinas, M., Athersuch, T.J., Font-Ribera, L., Villanueva, C.M., Vermeulen, R., Chadeau-Hyam, M., 2018. A multivariate approach to investigate the combined biological effects of multiple exposures. J Epidemiol Community Health 72, 564-571.

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software; Vol 1, Issue 5 (2008).

Le Cao, K.A., Boitard, S., Besse, P., 2011. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 12, 253.

Le Cao, K.A., Rossouw, D., Robert-Granie, C., Besse, P., 2008. A sparse PLS for variable selection when integrating omics data. Stat Appl Genet Mol Biol 7, Article 35.

Lenters, V., Portengen, L., Rignell-Hydbom, A., Jönsson, B.A.G., Lindh, C.H., Piersma, A.H., Toft, G., Bonde, J.P., Heederik, D., Rylander, L., Vermeulen, R., 2016. Prenatal Phthalate, Perfluoroalkyl Acid, and Organochlorine Exposures and Term Birth Weight in Three Birth Cohorts: Multi-Pollutant Models Based on Elastic Net Regression. Environmental Health Perspectives 124, 365-372.

Lenters, V., Portengen, L., Smit, L.A., Jonsson, B.A., Giwercman, A., Rylander, L., Lindh, C.H., Spano, M., Pedersen, H.S., Ludwicki, J.K., Chumak, L., Piersma, A.H., Toft, G., Bonde, J.P., Heederik, D., Vermeulen, R., 2015. Phthalates, perfluoroalkyl acids, metals and organochlorines and reproductive function: a multipollutant assessment in Greenlandic, Polish and Ukrainian men. Occup Environ Med 72, 385-393.

Lenters, V., Vermeulen, R., Portengen, L., 2018. Performance of variable selection methods for assessing the health effects of correlated exposures in case-control studies. Occup Environ Med 75, 522-529.

Lim, C., Yu, B., 2016. Estimation Stability With Cross-Validation (ESCV). Journal of Computational and Graphical Statistics 25, 464-492.

Louis, G.M., Weiner, J.M., Whitcomb, B.W., Sperrazza, R., Schisterman, E.F., Lobdell, D.T., Crickard, K., Greizerstein, H., Kostyniak, P.J., 2005. Environmental PCB exposure and risk of endometriosis. Hum Reprod 20, 279-285.

Manrai, A.K., Cui, Y., Bushel, P.R., Hall, M., Karakitsios, S., Mattingly, C.J., Ritchie, M., Schmitt, C., Sarigiannis, D.A., Thomas, D.C., Wishart, D., Balshaw, D.M., Patel, C.J., 2017. Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health. Annu Rev Public Health 38, 279-294.

Matta, K., Ploteau, S., Coumoul, X., Koual, M., Le Bizec, B., Antignac, J.P., Cano-Sancho, G., 2019. Associations between exposure to organochlorine chemicals and endometriosis in experimental studies: A systematic review protocol. Environ Int 124, 400-407.

Mostafavi, N., Jeong, A., Vlaanderen, J., Imboden, M., Vineis, P., Jarvis, D., Kogevinas, M., Probst-Hensch, N., Vermeulen, R., 2019. The mediating effect of immune markers on the association between ambient air pollution and adult-onset asthma. Sci Rep 9, 8818.

Mustieles, V., Fernandez, M.F., Martin-Olmedo, P., Gonzalez-Alzaga, B., Fontalba-Navas, A., Hauser, R., Olea, N., Arrebola, J.P., 2017. Human adipose tissue levels of persistent organic pollutants and metabolic syndrome components: Combining a cross-sectional with a 10-year longitudinal study using a multi-pollutant approach. Environ Int 104, 48-57.

Papadopoulou, E., Haug, L.S., Sakhi, A.K., Andrusaityte, S., Basagana, X., Brantsaeter, A.L., et al. 2019. Diet as a Source of Exposure to Environmental Contaminants for Pregnant Women and Children from Six European Countries. Environmental Health Perspectives 127: 107005.

Ploteau, S., Antignac, J.P., Volteau, C., Marchand, P., Vénisseau, A., Vacher, V., Le Bizec, B., 2016. Distribution of persistent organic pollutants in serum, omental, and parietal adipose tissue of French women with deep infiltrating endometriosis and circulating versus stored ratio as new marker of exposure. Environment International 97, 125-136.

Ploteau, S., Cano-Sancho, G., Volteau, C., Legrand, A., Vénisseau, A., Vacher, V., Marchand, P., Le Bizec, B., Antignac, J.-P., 2017. Associations between internal exposure levels of persistent organic pollutants in adipose tissue and deep infiltrating endometriosis with or without concurrent ovarian endometrioma. Environment International 108, 195-203.

Riffenburgh, R.H., 2006. Chapter 15 - Tests on Categorical Data, in: Riffenburgh, R.H. (Ed.), Statistics in Medicine (Second Edition). Academic Press, Burlington, pp. 241-279.

Roffman, D., Hart, G., Girardi, M., Ko, C.J., Deng, J., 2018. Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. 8, 1701.

Roy, A., Perkins, N.J., Buck Louis, G.M., 2012. Assessing Chemical Mixtures and Human Health: Use of Bayesian Belief Net Analysis. J Environ Prot (Irvine, Calif) 3, 462-468.

Sampson, J.A., 1927. Metastatic or Embolic Endometriosis, due to the Menstrual Dissemination of Endometrial Tissue into the Venous Circulation. The American journal of pathology 3, 93-110.143.

Schisterman, E.F., Perkins, N.J., Mumford, S.L., Ahrens, K.A., Mitchell, E.M., 2017. Collinearity and Causal Diagrams: A Lesson on the Importance of Model Specification. Epidemiology 28, 47-53.

Stafoggia, M., Breitner, S., Hampel, R., Basagana, X., 2017. Statistical Approaches to Address Multi-Pollutant Mixtures and Multiple Exposures: the State of the Science. Curr Environ Health Rep 4, 481-490.

Stingone, J.A., Pandey, O.P., Claudio, L., Pandey, G., 2017. Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among U.S. children. Environ Pollut 230, 730-740.

Sun, Z., Tao, Y., Li, S., Ferguson, K.K., Meeker, J.D., Park, S.K., Batterman, S.A., Mukherjee, B., 2013. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. Environ Health 12, 85.

Tape, T.G. 2001. Interpretation of Diagnostic Tests. Annals of Internal Medicine 135, 72.

Taylor, K.W., Joubert, B.R., Braun, J.M., Dilworth, C., Gennings, C., Hauser, R., Heindel, J.J., Rider, C.V., Webster, T.F., Carlin, D.J., 2016. Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology: Lessons from an Innovative Workshop. Environ Health Perspect 124, A227-a229.

Tomiazzi, J.S., Pereira, D.R., Judai, M.A., Antunes, P.A., Favareto, A.P.A., 2019. Performance of machine-learning algorithms to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke. 26, 6481-6491.

Upson, K., De Roos, A.J., Thompson, M.L., Sathyanarayana, S., Scholes, D., Barr, D.B., Holt, V.L., 2013. Organochlorine pesticides and risk of endometriosis: findings from a population-based case-control study. Environ Health Perspect 121, 1319-1324.

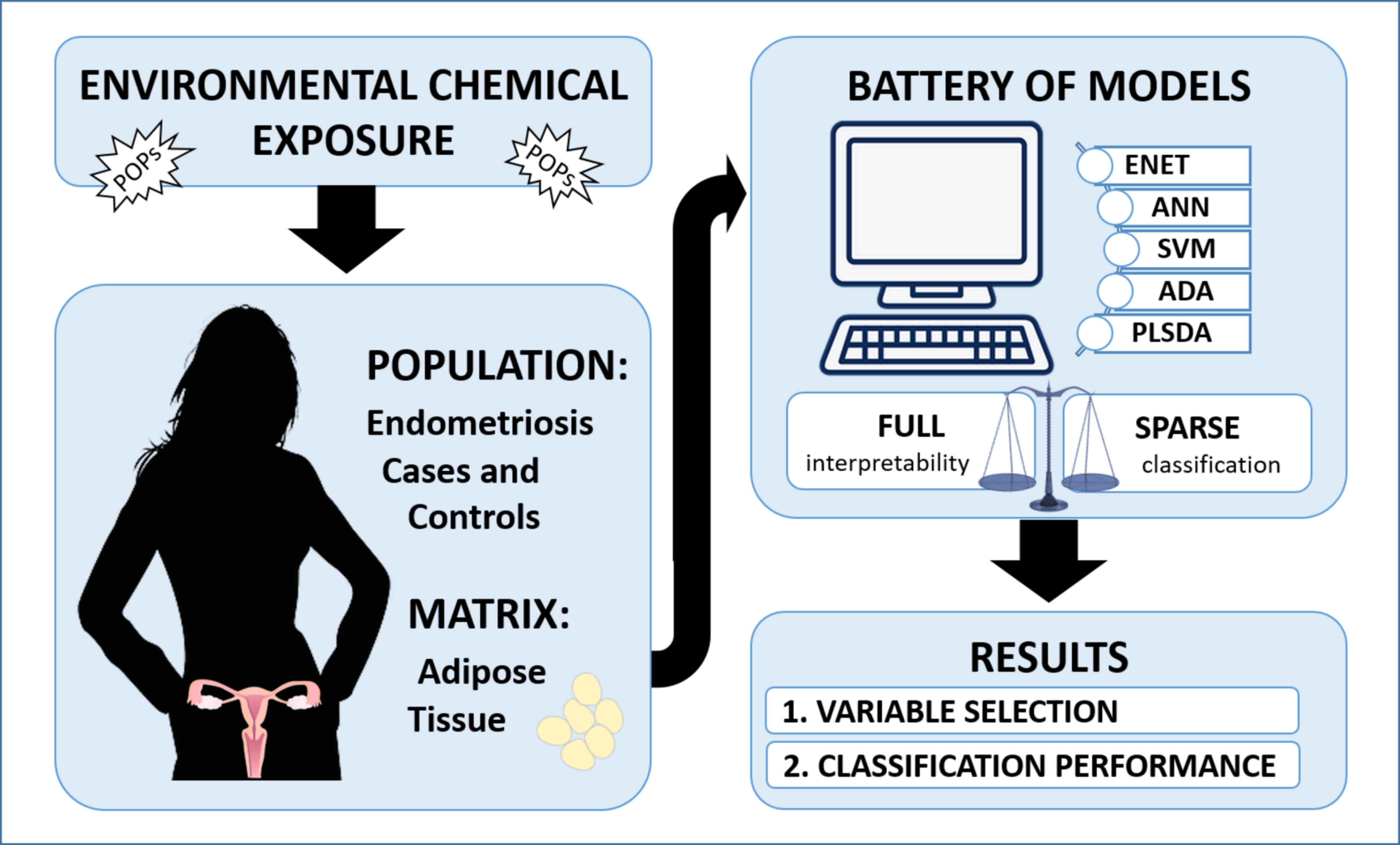
Vigneau, E., Chen, M., Qannari, E.M., 2015. ClustVarLV: An R Package for the Clustering of Variables Around Latent Variables.

Weisskopf, M.G., Seals, R.M., Webster, T.F., 2018. Bias Amplification in Epidemiologic Analysis of Exposure to Mixtures. Environ Health Perspect 126, 047003.

Zhang, B., Chen, Z., Albert, P.S., 2012. Latent class models for joint analysis of disease prevalence and high-dimensional semicontinuous biomarker data. Biostatistics 13, 74-88.

Zhao, Y., Li, J., Zhang, M., Lu, Y., Xie, H., Tian, Y., Qiu, W., 2019. Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study. Ear Hear 40, 690-699.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301-320.


ENVIRONMENTAL CHEMICAL EXPOSURE
POPs
POPs
POPULATION:
Endometriosis Cases and Controls
MATRIX:
Adipose Tissue
BATTERY OF MODELS
ENET
ANN
SVM
ADA
PLSDA
FULL
interpretability
SPARSE
classification
RESULTS
1. VARIABLE SELECTION
2. CLASSIFICATION PERFORMANCE