



HAL
open science

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

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

3

4

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

7

8 **Affiliations:**

9 ^a LABERCA, Oniris, INRAE, 44307, Nantes, France

10 ^b StatSC, ONIRIS, INRAE, Nantes, France

11 ^c Service de gynécologie-obstétrique, CIC FEA, Hôpital Mère Enfant, CHU Hôtel Dieu, Nantes,
12 France

13

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

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

17

18 **Keywords:**

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

21

22

23

24 **Abstract**

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

50

51 **Main findings capsule**

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

55

56

57 **Introduction**

58 Endometriosis is a hormone-dependent gynaecological disease characterised by the presence
59 of endometrial tissue outside the uterine cavity and contributes to a number of non-specific
60 symptoms, such as chronic pelvic pain, dysmenorrhea, dyschesia, dyspareunia, and often
61 infertility (Eskenazi et al., 2002; Giudice, 2010; Sampson, 1927). The precise aetiology of
62 endometriosis remains unclear but is likely multicausal, influenced by hormonal, genetic, and
63 environmental factors. Evidence from epidemiological studies suggests a relationship between
64 risk of endometriosis and exposure to some organochlorine persistent organic pollutants
65 (POPs) like dioxin 2,3,7,8-Tetrachlorodibenzodioxin (TCDD), polychlorobiphenyls (PCBs) and
66 organochlorine pesticides (OCPs) (Cano-Sancho et al., 2019), mechanistically supported by
67 experimental evidence (Bruner-Tran et al., 2010; Bruner-Tran and Osteen, 2010; Matta et al.,
68 2019). In a previous case-control study conducted in France (Ploteau et al., 2017), we found
69 statistically significant associations between presence of deep endometriosis and
70 concentrations of certain POPs in adipose tissue (AT), including 1,2,3,7,8-
71 pentachlorodibenzodioxin (PeCDD), octachlorodibenzofuran (OCDF),
72 polybromodiphenylether (PBDE) 183, polybromobiphenyl (PBB) 153 or cis-heptachlor
73 epoxide, among others. The approach previously used considered one pollutant at a time, with
74 multivariable logistic regression adjusting for known and suspected confounding variables.
75 This approach may prone to bias if associations are due to correlated coexposures. For this
76 reason, multipollutant models are today encouraged to evaluate coexposure-outcome
77 associations under collinear frameworks (Lenters et al., 2018; Weisskopf et al., 2018).
78 Collinearity is an acknowledged problem in analyses based on ordinary least squares (i.e.
79 linear regression). It occurs when two or more predictor variables are highly correlated, as is
80 often the case in datasets with mixtures of environmental chemical exposures. This may
81 exacerbate variances due to model misspecification, especially when prior biological
82 knowledge of underlying associations is not available (Schisterman et al., 2017). In the last
83 decades, novel statistical methods and computational frameworks have emerged motivated
84 by the challenges posed by air pollution mixtures, expanding the spectrum of available

85 approaches to address the various data constraints (Bellinger et al., 2017; Stafoggia et al.,
86 2017; Taylor et al., 2016). Overall, the models may be grouped by their capacities to reduce
87 data dimensionality, select variables (identify risk variables within highly redundant and
88 correlated variables) and group or cluster observations (Billionnet et al., 2012; Stafoggia et al.,
89 2017). Among epidemiological studies, however, application of multipollutant approaches
90 using biomarkers of exposure has been growing at a more modest pace. Some recent
91 simulation studies have compared the performance of several multipollutant models to identify
92 exposome-health associations with both continuous and dichotomous outcomes together with
93 their interactions (Agier et al., 2016; Barrera-Gomez et al., 2017; Lenters et al., 2018; Sun et
94 al., 2013). Results suggest there is no one-size-fits-all model and that model selection must
95 be made on the basis of the data structures.

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

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

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

119 **Methods and Materials**

120 *Study Population*

121 This study draws upon a case-control study conducted in Pays-de-la-Loire, France between
122 2013 and 2015, focusing on a group of 80 persistent pollutants analysed in the AT of a sample
123 population with and without endometriosis. Study design, recruitment, and methods have been
124 previously reported (Ploteau et al., 2017). Briefly, the study enrolled a total of 99 women ages
125 18-45. Cases (n= 55) included women diagnosed with deep endometriosis (with surgical
126 confirmation) and controls (n = 44) comprised a similar group of women present at the clinic
127 for other gynaecological issues unrelated to endometriosis, surgically confirmed to not have
128 endometriosis and displaying no related clinical symptoms (i.e. chronic pelvic pain,
129 dysmenorrhea, dyspareunia, infertility). From both groups, cases and controls, 2 g of parietal
130 AT (subcutaneous fat) samples were collected and stored at -80°C. Data were gathered
131 pertaining to the diagnosis, anthropometric variables, and other potentially relevant factors
132 such as age, body mass index (BMI), breastfeeding and parity. All participants signed an
133 informed consent form approved by the Bioethics Committee of GNEDS (Groupe Nantais
134 d'Éthique dans le Domaine de la Santé).

135 *Exposure Assessment*

136 Biomarkers of exposure were determined in adipose tissue, which is the most stable matrix for
137 POP measurements reflecting long-term exposure (Cano-Sancho et al., 2019). These
138 exposure estimates capture the window between onset of the first symptoms and the diagnosis
139 of endometriosis (7-10 years). The supporting methods used for chemical analyses have been

140 published elsewhere (Antignac et al., 2009; Bichon et al., 2015; Ploteau et al., 2016; Ploteau
141 et al., 2017). In brief, samples were quantified with ¹³C-labeled congeners using isotope
142 dilution and extracted under high temperature and pressure (ASE Dionex, Sunnyvale, CA,
143 USA). Gravimetric methods were used to measure fat content, and extracts were reconstituted
144 in hexane for cleanup. OCPs were isolated using gel permeation chromatography; other target
145 substances were isolated using three successive purification steps: acid silica, Florisil®, and
146 celite/carbon columns. PCDD/F, PCB, PBDE, PBB and OCP were measured by gas
147 chromatography (Agilent 7890A) coupled with high-resolution mass spectrometry (GC-HRMS)
148 on double sector instruments (JEOL MS 700D and 800D) after electron impact ionization (70
149 eV), operating at 10000 resolutions (10% valley) and in the single ion monitoring (SIM)
150 acquisition mode. HBCD isomers were quantified using liquid chromatography coupled with
151 tandem mass spectrometry (LC-MS/MS) on a triple quadrupole instrument (Agilent 6410) using
152 electrospray ionization and selective reaction monitoring. The full list of analysed chemicals
153 and congeners can be found in the Supplemental Table S1. All methods were validated
154 according to Regulation (EU) No 376/2014 of the European Parliament (EU, 2014). Analysis
155 was performed in an ISO 17025:2005 accredited laboratory. All internal exposure data were
156 generated blinded to the case/control status of samples. Recoveries were in the 80–120%
157 range, and expanded uncertainty was lower than 20%. Exposure levels for POPs were
158 expressed in a lipid-weight basis (lw).

159 *Data pre-processing*

160 Missing data were characterised to determine their nature (Missing at Random vs Missing Not
161 at Random). Numerical covariates missing at random (i.e. BMI, age) were imputed using MICE
162 package in R. Distributions before and after imputation were checked to ensure consistency.
163 Data missing not at random included several POPs that were either not detected through
164 quantification or were found to be below the limit of detection. Exposure variables lower than
165 the limit of detection (LOD) were assigned a value of LOD/2 (Cohen and Ryan, 1989).
166 Variables for which over 75% of exposure data were missing or below LOD were excluded

167 from analysis for quality control purposes (See Table S1). Remaining exposure variables were
168 log transformed, centred and scaled by their standard deviations.

169 *Exploratory Data Analysis*

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

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

180 *Multipollutant Data Analysis*

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

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

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

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

205 We also computed metrics for classification performance including Receiver Operating
206 Characteristic (ROC), Area Under the Curve (AUC), sensitivity, and specificity. ROC curves
207 measure a test's ability to discriminate between cases and controls and is quantified by the
208 AUC. An AUC of 1 means the test has 100% discriminative capacity, and a value of 0.5 means
209 the test is unable to discern cases from controls any more than random chance. In general,
210 values between 0.9-1.0 are considered very good, values 0.8-0.9 are considered good, 0.7-
211 0.8 as fair, 0.6-0.7 as poor, and 0.5-0.6 as failure (Tape, 2001). Sensitivity measures the
212 capacity of the model to correctly identify positive cases, while specificity indicates the capacity
213 to correctly identify controls. McNemar's test on paired proportions was used to assess the
214 predictive accuracy of the classification model. We also compared the agreement of variables
215 selected between models, their VI, their interpretability and flexibility to be applied in
216 epidemiological studies.

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

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

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

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

234 b) Artificial Neural Network

235 ANNs are ML algorithms inspired by the structure of biological neural networks. They consist
236 of a number of interconnected neural nodes. The structure of ANNs usually comprise three
237 principal layers: the input layer includes input nodes (predictor variables), the output layer
238 consists of a single output node (endometriosis status), and the middle hidden layer(s) are
239 populated by a collection of hidden nodes with values which model the complex relationships
240 between the input and output layers but which do not of themselves have a real world
241 analogue. The synapses which connect each of these layers' nodes to one another are
242 weighted, which represents the strength of the connection, similar to coefficients in logistic
243 regression models. In neural networks, the weight decay value acts as the regularisation term.
244 ANN is implemented using the *nnet* package. Tuning parameters are *size* (number of hidden
245 layers) and *decay* (weight decay). Size ranged from 1 to 50, and decay from 0 to 0.9.

246 c) Support Vector Machine

247 SVM is a classifier that works by reimagining data in a multidimensional space and generating
248 multiple potential hyperplanes to separate data, then selecting the optimal hyperplane which

249 maximises the margins between the two groups (here, cases and controls). Typically, SVMs
250 are used as a linear classification model, but they can also generate hyperplanes for nonlinear
251 data using a kernel function. In this study, we used an SVM with a radial basis function (RBF)
252 kernel for nonlinear data to transform the original feature space for better separation of the two
253 groups. Regularisation is controlled by a cost parameter. The cost parameter controls the
254 tradeoff between training errors and model complexity. A smaller cost value increases the
255 number of training errors while larger costs may lead to overfitting. The sigma parameter with
256 RBF kernel determines the flexibility of the decision boundary and how much influence a single
257 feature can exert. Larger sigmas create a more flexible and smooth decision boundary with
258 more variance and thus act as a more general classifier, while smaller sigma values are stricter
259 and tend to make more local classifiers (Ben-Hur A., 2010). Tuning parameters of *svmradiial*
260 from the *kernelab* package are *sigma* (Sigma) and *C* (Cost). Sigma ranged from 0.001 to 1, and
261 cost ranged from 0 to 100.

262 d) Boosting trees: Adaboost

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

273 e) Partial least squares discriminant analysis

274 Partial least squares discriminant analysis (PLSDA) models approximate the relationship
275 between predictor variables and the response variable (endometriosis status), searching for
276 directions of maximum covariance between the two. Using the *softmax* function, predictor

277 variables are assigned "probability-like" values (on a scale of 0 to 1 which sum to 1), and the
278 class with the largest class probability is the predicted class. In the sparse form, only the most
279 predictive or discriminative features from the data are selected to inform classification.

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

282

283 **Results**

284 *Descriptive analysis*

285 Cases (n = 55) and controls (n = 44) were matched for age, BMI, and breastfeeding history,
286 three factors which are known to be strongly correlated with internal exposure levels of POPs
287 (Ploteau et al., 2016). Mean and standard deviation age of control and case group were 32.6
288 (± 6.5) and 34.3 (± 6.2) years, respectively (Student T test, $p=0.19$). BMI also did not differ
289 between groups, with 25.4 (± 5.9) kg/m² for controls and 24.0 (± 5.1) kg/m² for cases ($p=0.21$).
290 Parity and breastfeeding were not included in the models due to their uncertain causal role in
291 the pathogenesis of endometriosis (Ploteau et al., 2017; Upson et al., 2013). Cases exhibited
292 lower average breastfeeding duration (4.1 \pm 14.9 months) than controls (1.3 \pm 3.1 months), but
293 did not differ statistically ($p=0.18$). Distributions of concentrations of POPs in AT for cases and
294 controls are provided in Supplemental Table S2.

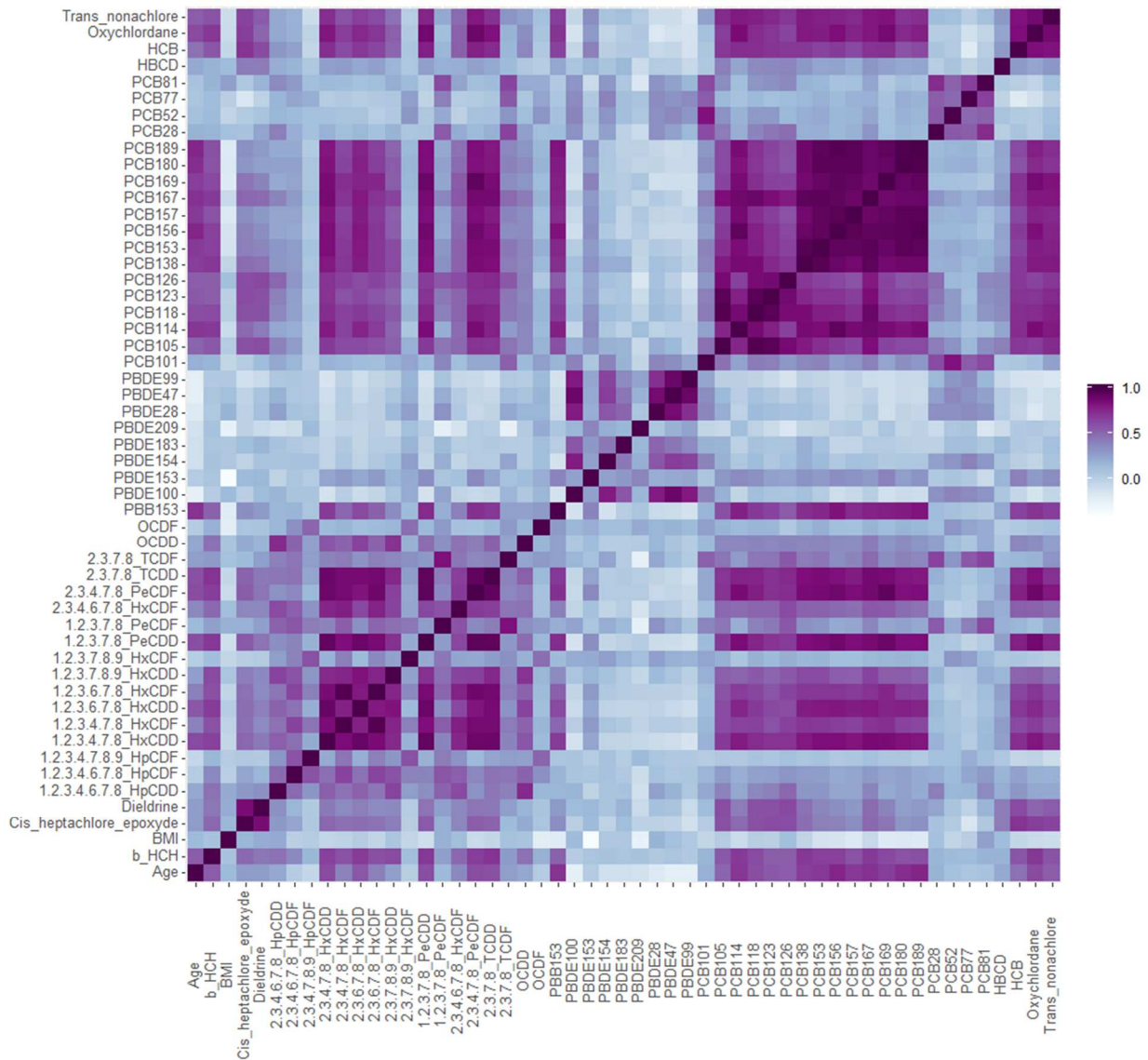
295

296 *Exploratory Data Analysis*

297 Coefficients from the bivariate correlation analysis between pollutants are depicted in the
298 heatmap in Figure 1. Clusters of dioxins, PCBs, brominated flame retardants and pesticides
299 present positive correlations. Coplanar PCBs 189, 169, 167, 157, 156, 126, 123, 118, 114, 105
300 were found to be positively correlated amongst one another but not with coplanar PCBs 77
301 and 81. Interestingly, OCDF was not found to be strongly correlated with any other variable,
302 save for a moderate positive association with 1.2.3.7.8.9 HxCDF and 1.2.3.4.7.8.9 HpCDF.
303 PBDEs were found to be mildly negatively correlated with dioxins, furans, and pesticides,
304 1.2.3.7.8 PeCDF and 2.3.7.8 TCDF, and non-coplanar PCBs 28, 52, and 101. Age was mildly

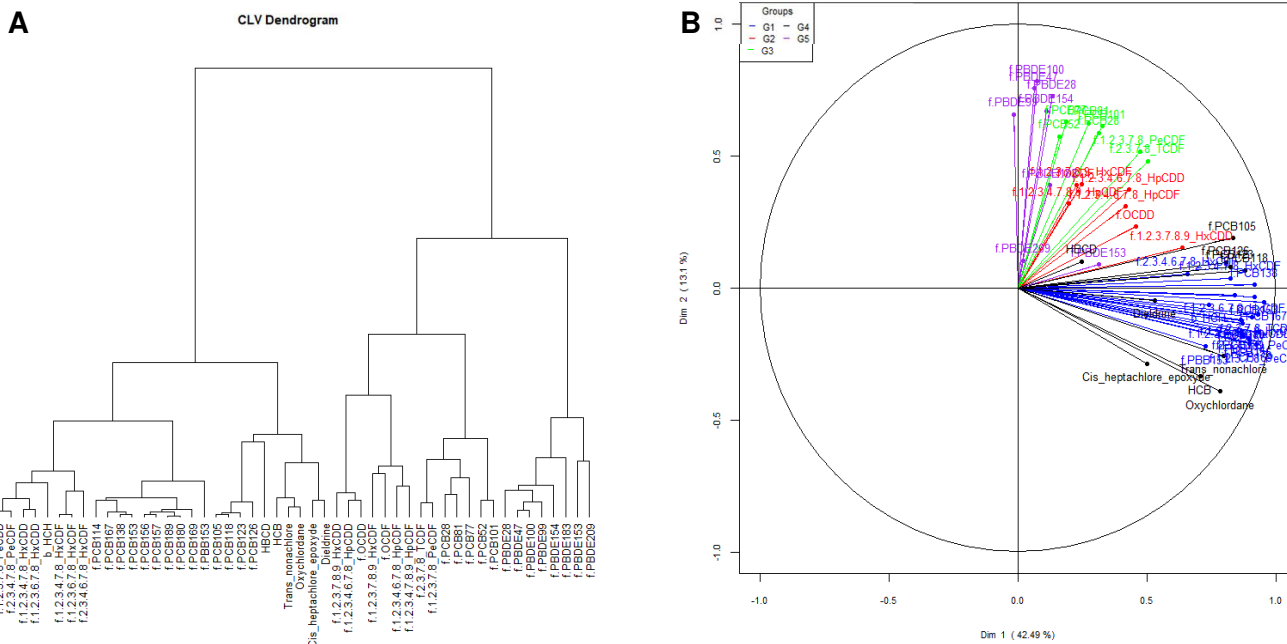
305 positively correlated with the same clusters of dioxins and coplanar PCBs. BMI did not show
 306 any correlations with any of the other variables. Heatmaps displaying the correlation analysis
 307 stratified by endometriosis status did not show visual differences between cases and controls
 308 (Supplemental Figure S1).

309 **Figure 1.** Correlation analysis heatmap



310
 311
 312 With regard to PCA, the two first components summarise more than a half of the data (42.49%
 313 of inertia retrieved by the first component, 13.10% by the second) (Supplemental Figure S2A).
 314 Factor maps depicting the correlations between pollutants variables and the two components
 315 are available in Supplemental Figures S2B-C.

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



319
 320
 321 CLV revealed the underlying structure of the data, which can be visualised in a dendrogram
 322 (Figure 2A), five clusters (K = 5) of which can be seen in a two dimensional PCA variables
 323 configuration (Figure 2B). The groups identified tend to form clusters around extant chemical
 324 families: dioxins, furans, pesticides, coplanar PCBs, non-coplanar PCBs, PBDEs, and PBBs.
 325 The partition of variables has been defined so that within each cluster the angles between
 326 vectors associated with the exposure variables and a latent (not observed) central variable are
 327 minimised (maximising correlation). However, some exposure variables such as HCB (G4),
 328 or PBDE209 and PBDE153 (G5) which are both far from the centre of their respective cluster
 329 and not well represented into the first PCA map may have been difficult to assign to any of the
 330 five clusters highlighted. In the dendrogram, it can be seen that HCB would be in its own
 331 cluster at K = 11, and that PBDE209 and PBDE153 form a very small cluster.

332 *Multipollutant Data Analysis*

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

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

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

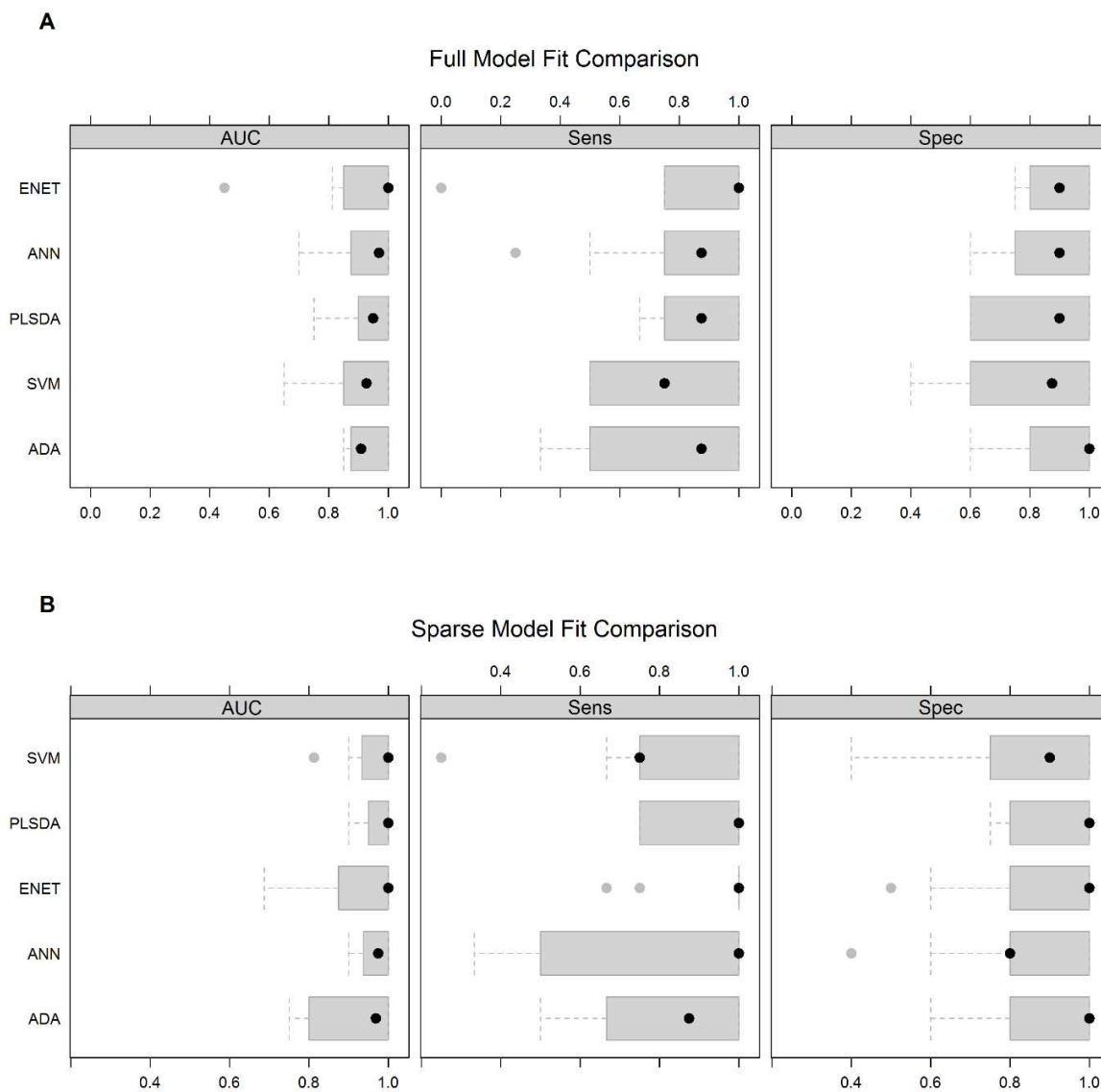
337

338 **Full Models**

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

341 Models were further compared according to their fit (Figure 3A, Table S3) and classification
 342 performance (Table S4) using a confusion matrix to determine accuracy, AUC, sensitivity, and
 343 specificity. Ridge, SVM and ANN scored highest in AUC (SD) (0.968 (0.035), 0.958 (0.059),
 344 0.956 (0.063) respectively. ENET had the highest scoring sensitivity (0.900 (0.129)) and ANN
 345 had the highest scoring specificity (0.900 (0.175)) with the lowest sensitivity (0.775 (0.208)).

346 **Figure 3.** Model Fit Comparison for (A) Full and (B) Sparse Models presented in median and
 347 interquartile range.
 348



349
 350 **Sparse Models**
 351 Calibration plots of variable selection for each model are available in Supplemental Figures S9-
 352 S13. Nineteen variables were identified by ENET (OCDF, cis-heptachlor epoxide, PCB77,
 353 PCB81, BMI, PCB123, trans-nonachlor, PCB52, PCB101, PCB157, 2,3,4,6,7,8 HxCDF, PBB153,
 354 1,2,3,4,6,7,8 HpCDF, Oxychlordane, PBDE183, PBDE154, and 1,2,3,4,6,7,8 HpCDD); twenty

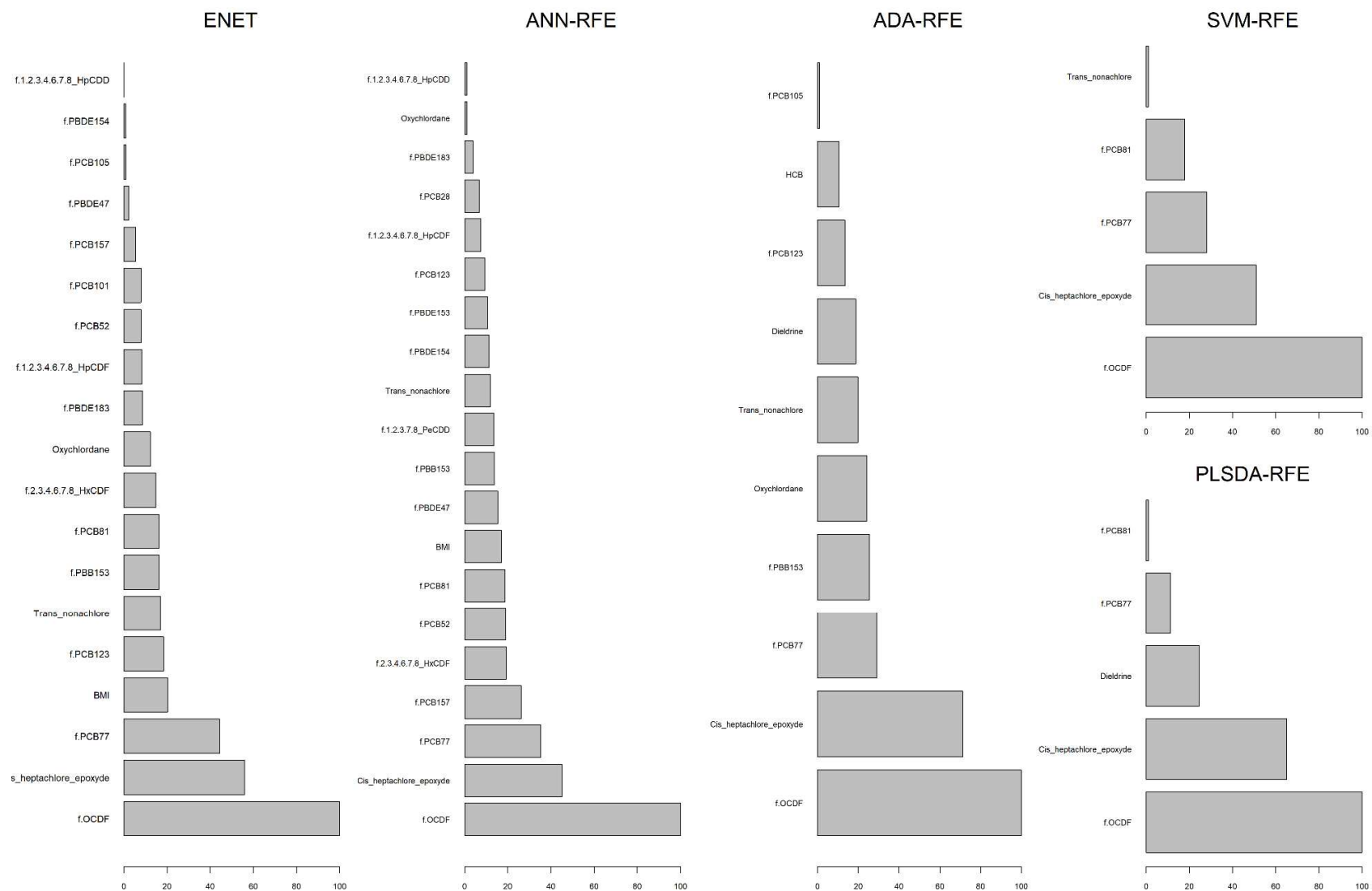
355 variables were identified by ANN (OCDF, cis-heptachlor epoxide, PCB77, PCB81, PBB153, BMI,
356 2.3.4.6.7.8 HxCDF, PCB157, 1.2.3.7.8 PeCDD, PBDE154, PBDE47, PCB52, trans-nonachlor,
357 PCB28, PBDE153, PCB123, 1.2.3.4.6.7.8 HpCDD, oxychlorane, PBDE183, 1.2.3.4.6.7.8
358 HpCDF); five were identified by SVM (OCDF, cis-heptachlor epoxide, PCB77, PCB81, trans-
359 nonachlor); ten were identified by ADA (OCDF, cis-heptachlor epoxide, PCB77, PBB153,
360 oxychlorane, trans-nonachlor, dieldrin, PCB123, HCB, PCB105) and five by PLSDA (OCDF, cis-
361 heptachlor epoxide, dieldrin, PCB77, and PCB81) (Figure 4).

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

365 Summary of classification performance metrics (accuracy, AUC, sensitivity, and specificity) are
366 presented in Figure 4B and Table S4 for each model. Model fit accuracy for sparse models all
367 ranged from 85.0-88.8%, and AUC indices (SD) were all greater than 0.95 (ENET 0.988 (0.024),
368 ANN 0.989 (0.024), SVM 0.973 (0.058), ADA 0.954 (0.039), PLSDA 0.980 (0.045)). Sensitivity
369 across models did not vary markedly from one another (ENET 0.817 (0.211), ANN 0.900 (0.129),
370 SVM 0.891 (0.142), ADA, 0.867 (0.188), PLSDA 0.975 (0.079)), nor did specificity (ENET 0.915
371 (0.111), ANN 0.895 (0.146), SVM 0.885 (0.256), ADA, 0.870 (0.106), PLSDA 0.775 (0.203)).
372 Values of all model fit metrics are available in Supplemental Table S3.

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

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



379

380

381 **Discussion**

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

401 This initial exploration supports the use of regularised regression (i.e. elastic-net) for variable
402 selection, exhibiting an adequate balance between classification performance and interpretability.
403 In this study, powerful classifiers such as SVM, ANN or ADA did not outperform other commonly
404 used algorithms such as PLSDA or ENET. Globally, variable selection was very consistent across

405 the different models with minor differences in the biomarker rankings. For sparse models, the
406 number of discriminant variables retained was substantially lower for SVM and PLSDA than for
407 the other models. Three variables appeared as the strongest predictors of endometriosis status,
408 namely OCDF, cis-heptachlor epoxide, and PCB77. Trans-nonachlor and PCB81 were identified
409 by four of the five tested models, while PBB153, PCB123, and oxychlorane were identified by
410 three. These results are consistent with our previous findings using a sequential logistic
411 regression followed by false discovery rate correction, with an Odds Ratio (95% CI) of 5.42 (2.73-
412 12.85) and 5.36 (2.44-14.84) for OCDF and cis-heptachlor epoxide, respectively (Ploteau et al.,
413 2017). Coplanar PCB 77 and 123, as well as polybrominated flame retardant PBB153, were also
414 identified as important predictors. The correlations between pesticides cis-heptachlor epoxide and
415 trans-nonachlor with PCDDs, coplanar PCBs, several furans and non-coplanar dioxins might
416 mask the impact of the latter on endometriosis in sparse models. Interestingly, OCDF, the
417 strongest signal identified by all five models, was not strongly correlated with any other predictor
418 variable.

419 The model fit of five models did not differ substantially in terms of AUC, specificity, or sensitivity.
420 In this study, all five models had AUC values greater than 0.9, suggesting that this battery of
421 algorithms presents a promising method of modelling the associations between concentrations of
422 POPs in AT and endometriosis status. Interestingly, PLSDA with RFE performed well in model fit
423 (AUC = 0.98) but scored lowest in classification accuracy (i.e. 0.68 (95% CI; 0.43, 0.87)). This
424 may be due to the use of RFE to induce sparsity, instead of using the intrinsic sparse PLSDA
425 (sPLDSA) with lasso penalisation of PLS loading vectors (Le Cao et al., 2011; Le Cao et al.,
426 2008). The performance of the multiclass wrapper RFE has shown to decrease dramatically with
427 the number and correlation of variables due to the backward elimination used for variable
428 selection (Le Cao et al., 2011). We have applied RFE here to allow direct comparison among
429 models; however, future studies with wide and highly correlated datasets should consider the use

430 of sPLSDA over the RFE procedure. Surprisingly, powerful classifiers such as ANN and SVM did
431 not outperform the classification performance of more standard methods such as ENET with the
432 present dataset. The small sample size of the dataset might explain the imperfect architecture of
433 hidden layers, the number of neurons in each layer, and the activation functions in ANN
434 (Alwosheel et al., 2018).

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

455 control group but failed to differentiate between five different exposure classes (Tomiazzi et al.,
456 2019).

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

474 As the variable selection process should be considered a preliminary step previous to inferential
475 analysis, an additional challenge posed by ML is the interpretability of outputs. ML algorithms are
476 often viewed as “black boxes,” where it is difficult to inspect the inner workings of how outputs are
477 generated and what they mean in a real-world context. The coupling of modelling techniques with
478 graphical approaches has been proposed as a crucial way to apply and interpret ANNs in
479 epidemiological research (Duh et al., 1998). In a simulation setting, kernel mapping in combination

480 with a perceptron neural network has shown to efficiently generate odds ratios from perceptron
481 weights to ease epidemiological interpretation of complex nonlinear exposure-disease
482 associations (Heine et al., 2011). Future simulation studies should aim to extend the knowledge
483 of model performance of ML classifiers in exposome-health settings, exploring the impact of
484 parametrisation, sample size, correlation and interaction between exposure variables (Barrera-
485 Gomez et al., 2017; Lenters et al., 2018).

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

505 **Conclusions**

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

518

519 **Declaration of Interest**

520 Authors declare no conflicts of interest.

521

522 **Acknowledgements**

523 KM received a French regional government grant for doctoral allocations Pays de la Loire CPER
524 2014-2020.

525

526 **References**

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

530 Statistical Methods to Assess Exposome-Health Associations. *Environ Health Perspect* 124,
531 1848-1856.

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

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

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

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

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

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

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

553 Bichon, E., Guiffard, I., Venisseau, A., Marchand, P., Antignac, J.P., Le Bizec, B., 2015. Ultra-
554 trace quantification method for chlordecone in human fluids and tissues. *J Chromatogr A* 1408,
555 169-177.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

604 Lenters, V., Portengen, L., Rignell-Hydbom, A., Jönsson, B.A.G., Lindh, C.H., Piersma, A.H., Toft,
605 G., Bonde, J.P., Heederik, D., Rylander, L., Vermeulen, R., 2016. Prenatal Phthalate,
606 Perfluoroalkyl Acid, and Organochlorine Exposures and Term Birth Weight in Three Birth Cohorts:
607 Multi-Pollutant Models Based on Elastic Net Regression. Environmental Health Perspectives 124,
608 365-372.

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

614 Lenters, V., Vermeulen, R., Portengen, L., 2018. Performance of variable selection methods for
615 assessing the health effects of correlated exposures in case-control studies. Occup Environ Med
616 75, 522-529.

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

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

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

626 Matta, K., Ploteau, S., Coumoul, X., Koual, M., Le Bizec, B., Antignac, J.P., Cano-Sancho, G.,
627 2019. Associations between exposure to organochlorine chemicals and endometriosis in
628 experimental studies: A systematic review protocol. *Environ Int* 124, 400-407.

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

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

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

639 Ploteau, S., Antignac, J.P., Volteau, C., Marchand, P., Vénisseau, A., Vacher, V., Le Bizec, B.,
640 2016. Distribution of persistent organic pollutants in serum, omental, and parietal adipose tissue
641 of French women with deep infiltrating endometriosis and circulating versus stored ratio as new
642 marker of exposure. *Environment International* 97, 125-136.

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

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

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

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

653 Sampson, J.A., 1927. Metastatic or Embolic Endometriosis, due to the Menstrual Dissemination
654 of Endometrial Tissue into the Venous Circulation. *The American journal of pathology* 3, 93-
655 110.143.

656 Schisterman, E.F., Perkins, N.J., Mumford, S.L., Ahrens, K.A., Mitchell, E.M., 2017. Collinearity
657 and Causal Diagrams: A Lesson on the Importance of Model Specification. *Epidemiology* 28, 47-
658 53.

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

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

665 Sun, Z., Tao, Y., Li, S., Ferguson, K.K., Meeker, J.D., Park, S.K., Batterman, S.A., Mukherjee, B.,
666 2013. Statistical strategies for constructing health risk models with multiple pollutants and their
667 interactions: possible choices and comparisons. *Environ Health* 12, 85.

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

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

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

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

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

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

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

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

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

690

691

ENVIRONMENTAL CHEMICAL EXPOSURE



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