Statistical methods for studying mixtures and the exposome

Alan Domínguez, Maximilien Génard-Walton, Charline Warembourg

With materials from Augusto Anguita & Xavier Basagaña

ISGlobal Barcelona Institute for Global Health

iABD EPIGENETICS CHRONIC DISEASES CANCER

ISEE Young Rennes
Pre-conference workshop – June 7, 2024

Inserm

irset Institut de recherche environnement santé travail

Outline

- Packages installation
- Theorical presentation
  - 1. Introduction
  - 2. Statistical analysis
    - 2.1 ExWAS and variable selection
    - 2.2 Cluster analysis
    - 2.3 Mixture models
  - 3. Conclusion
- Hands-on session

image: freepik.com

Package installation

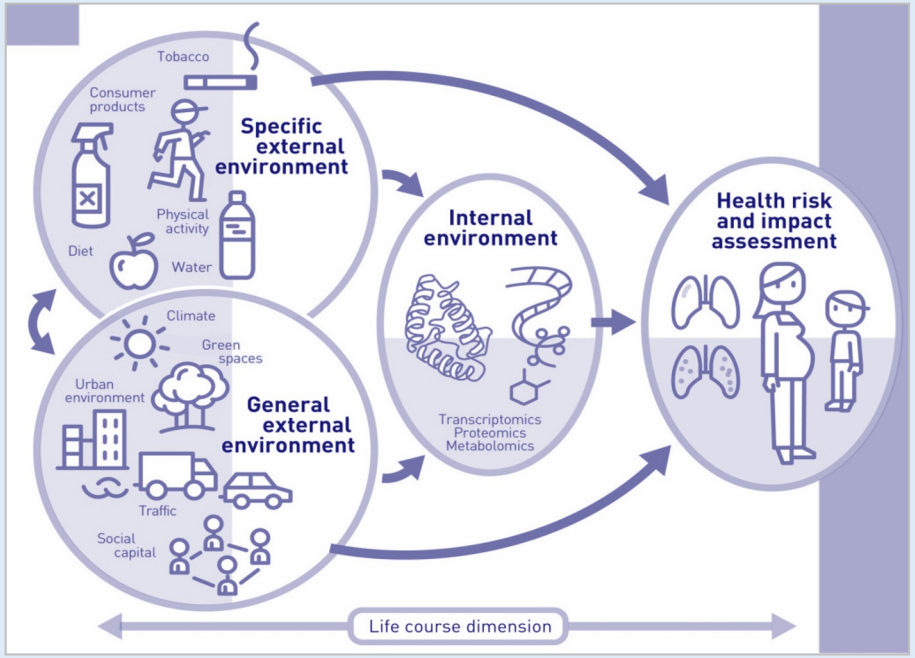
https://github.com/alldominguez/isee_young_rennes_ws1

1.
Introduction

The Exposome

*"The **exposome** encompasses the totality of human environmental (i.e., non-genetic) exposures from conception onwards, complementing the genome".*

(Chris Wild 2005)



The mixtures

Mostly used to study multiple exposure to chemicals but applicable to other exposures

→ Combined effect (additive or synergic)

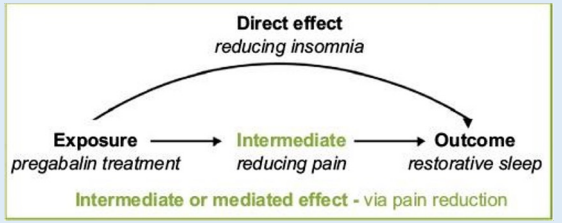
https://storyset.com/medical

Main difference between exposome and mixture analysis = **Data dimension**

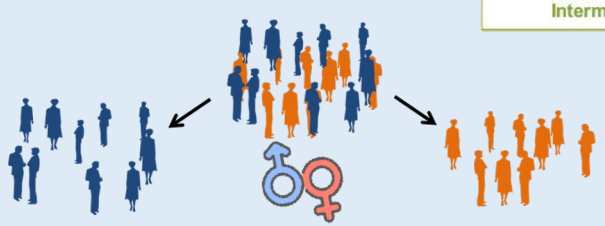
Classical analysis strategy in epidemiology studies

- The **research question** is often well defined:
    - How exposure X affects Y, controlling for possible confounding factors C1,..., C_P ?

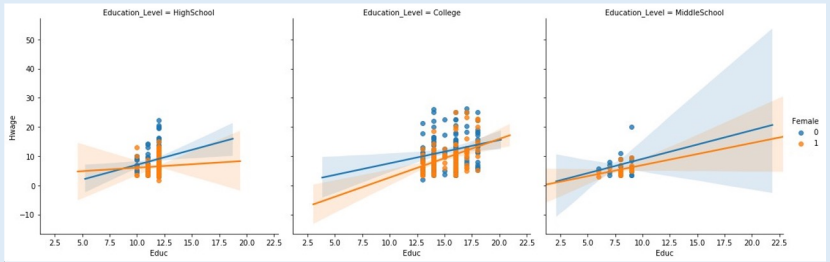
- Direct effect, indirect, % of mediating effect



- Stratified analysis (subgroups)

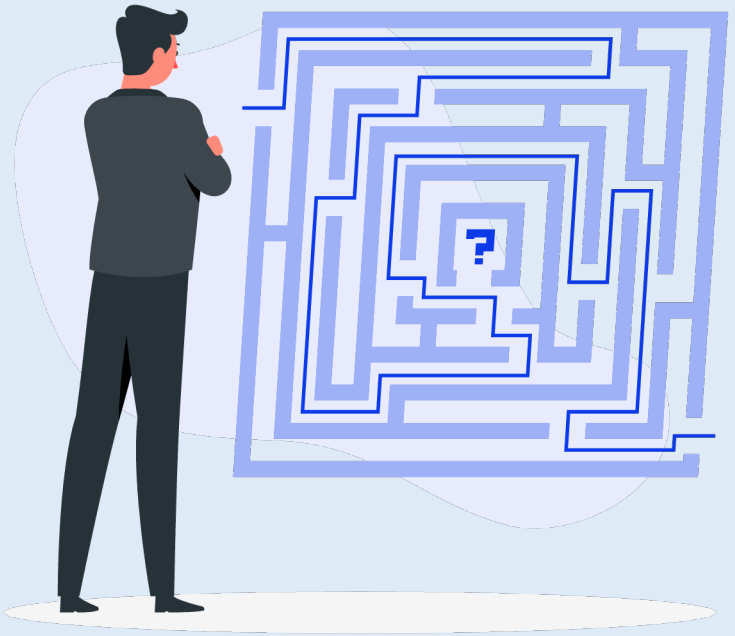


- Interactions between exposure and others factors



But in the case of mixture and the exposome…

- We will have to face **challenges such** as:
  - High-dimension and correlated data
  - Missing values
  - Measurement error
  - Interactions
  - Non-linearity
  - Potential bias (selection, confusion, etc.)
  - Repeated data



https://storyset.com/people

- New statistical methods have been developed to address (part of) these issues

Main issues in exposome analysis

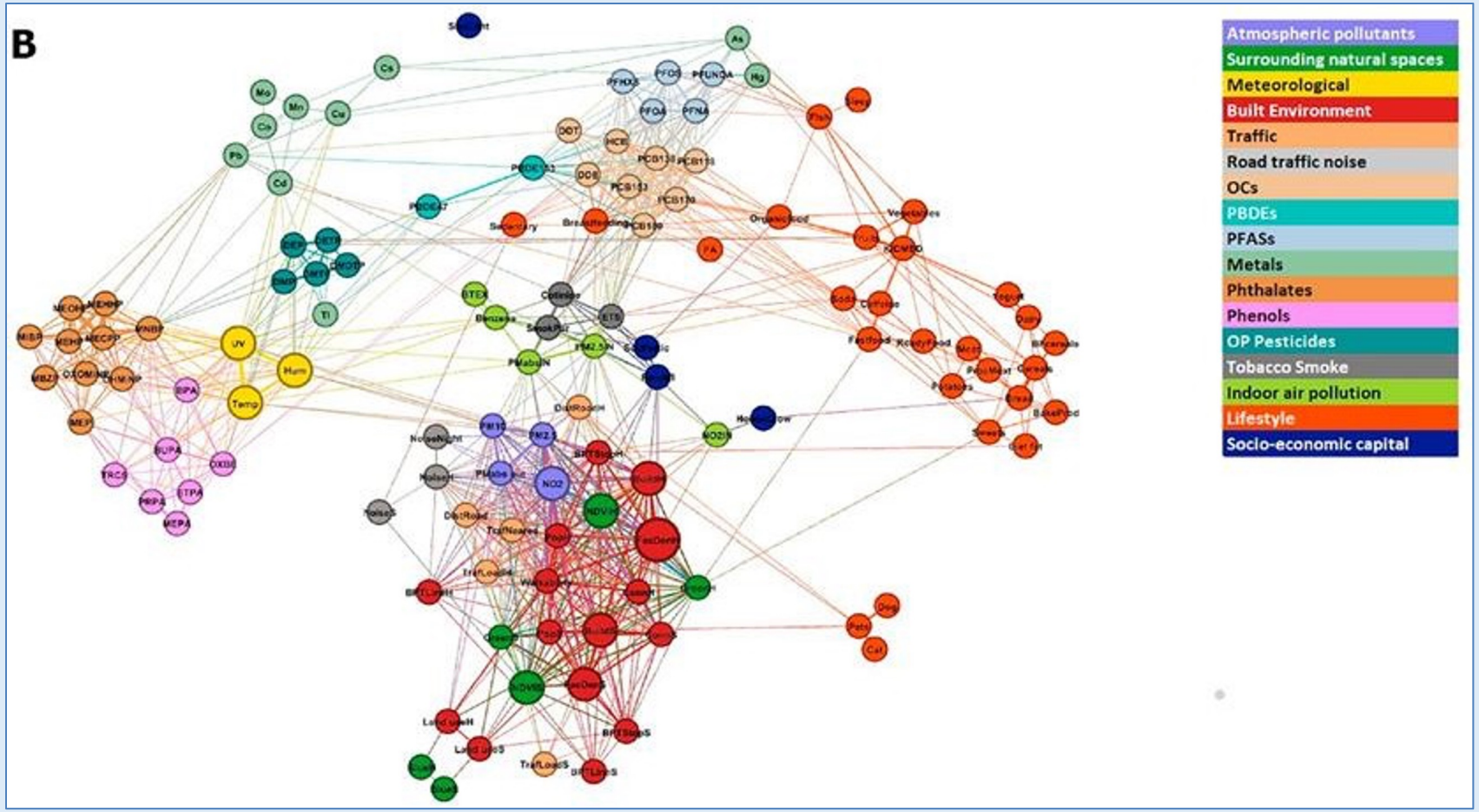
- High dimensionality and correlated variables

Helix subcohort

n=1301 children

Postnatal period: 6-11 years old

122 exposure variables



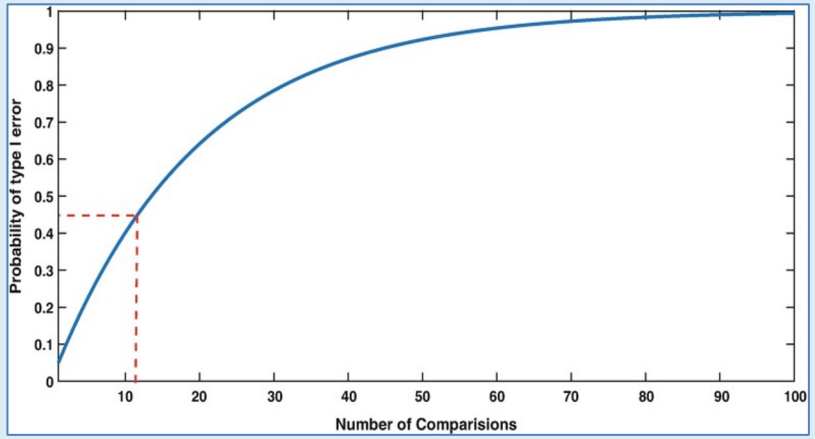
Tamayo-Uria, Environ Int 2019

Issues arising from multiple testing

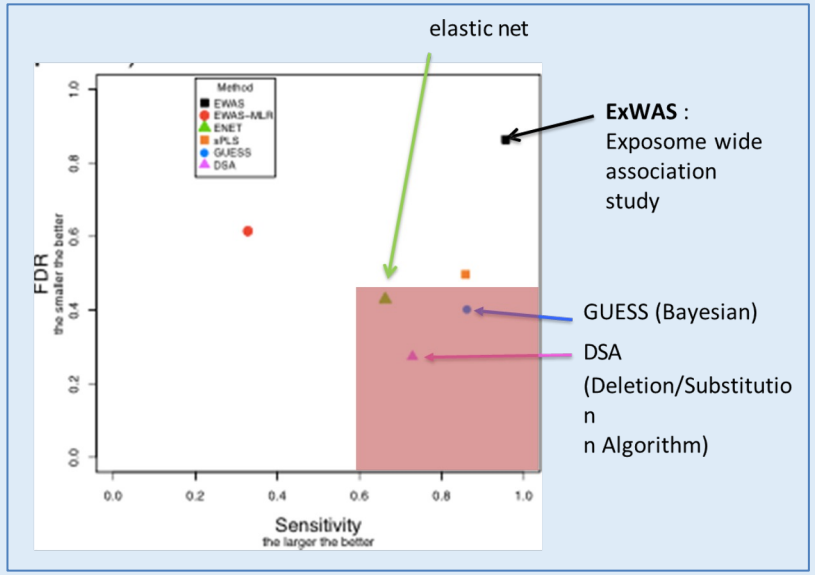
- The risk of false positive increases as the number of exposures of interest increases

- Correction methods:
  - **Family-wise error correction** (FWER): controls for the probability of at least one false discovery
    e.g., Bonferroni
  - **False discovery rate** (FDR): controls the proportion of false discoveries
    e.g., Benjamini & Hochberg

- We need methods that **limit the false discovery rate** but also **preserve the risk of false negatives**



Herzog 2019

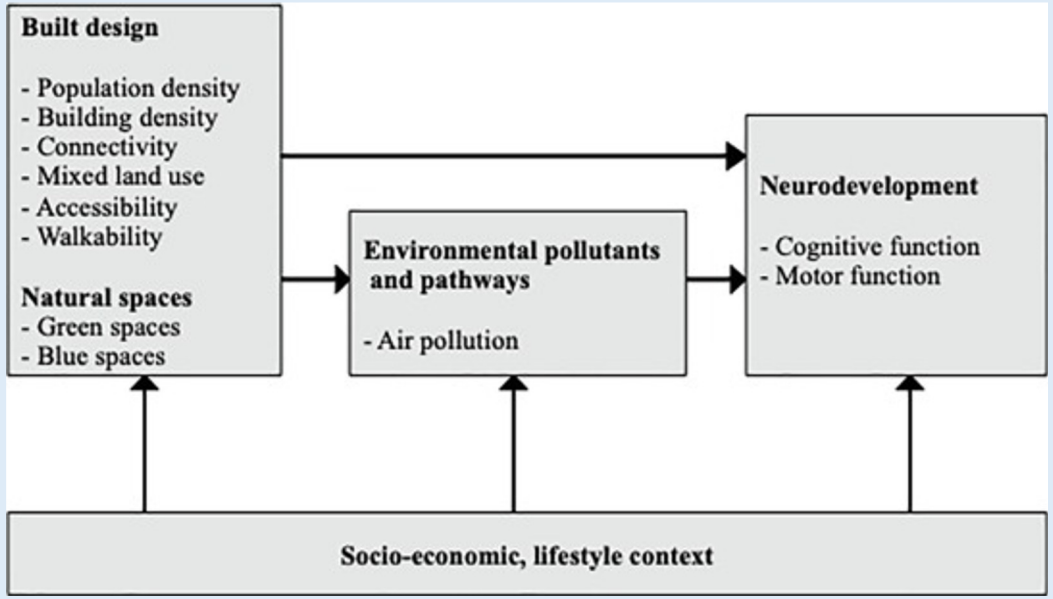


Agier EHP 2016

Confusion in multiple exposures context

- **Complex confounding structures** with multiple relationships existing between factors

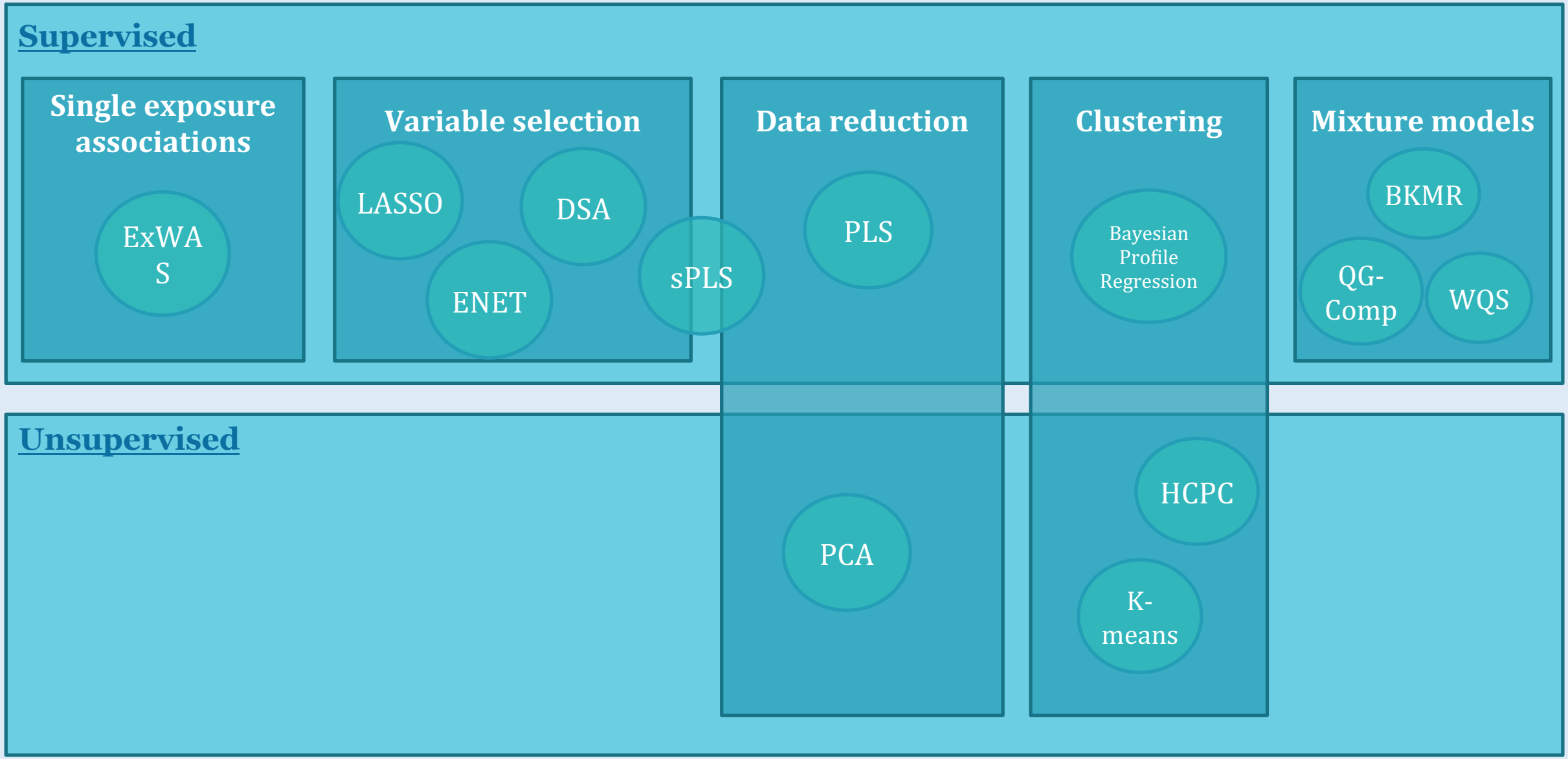
- Classical regression models are not are able (or little efficient) to accommodate high number of (correlated) variables



Binter Environ Int 2022

- Risk of **overadjustment** or **residual confounding** (same set of confounders for all exposures?)

Sample of available methods to deal with multiple variables



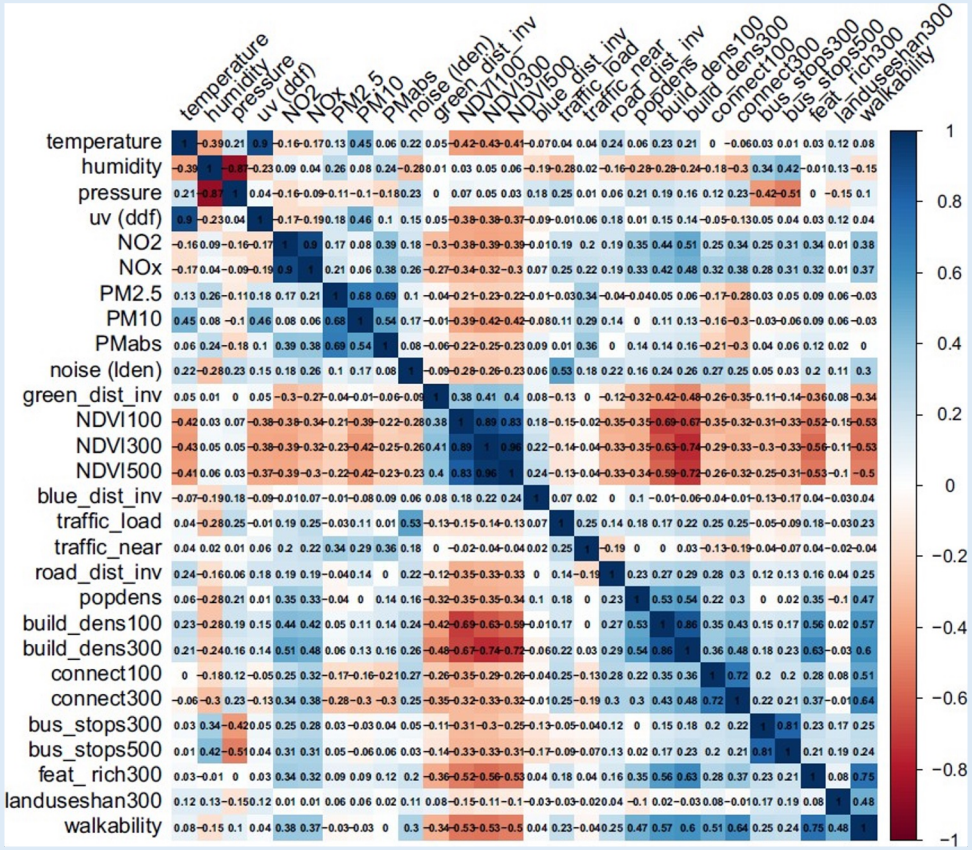
Exposome analysis

- **There are different potential research questions**, that will lead us to choose different approach, e.g.,:

1. **Are there one or more exposures associated with the outcome?**
2. **Are there specific exposure profiles associated with the outcome?**
3. **Are there mixtures of exposures that affect the outcome?**

Descriptive analysis

- **Fundamental part** of any analysis to become familiar with data
  - Description of variables (+ by subgroups)
  - Variable distributions → transformation?
  - Missing data → imputation?
  - Correlation between variables



Correlation heatmap - *Robinson, EHP 2018*

2.1

Association analysis

Are there one or more exposures associated with the outcome?

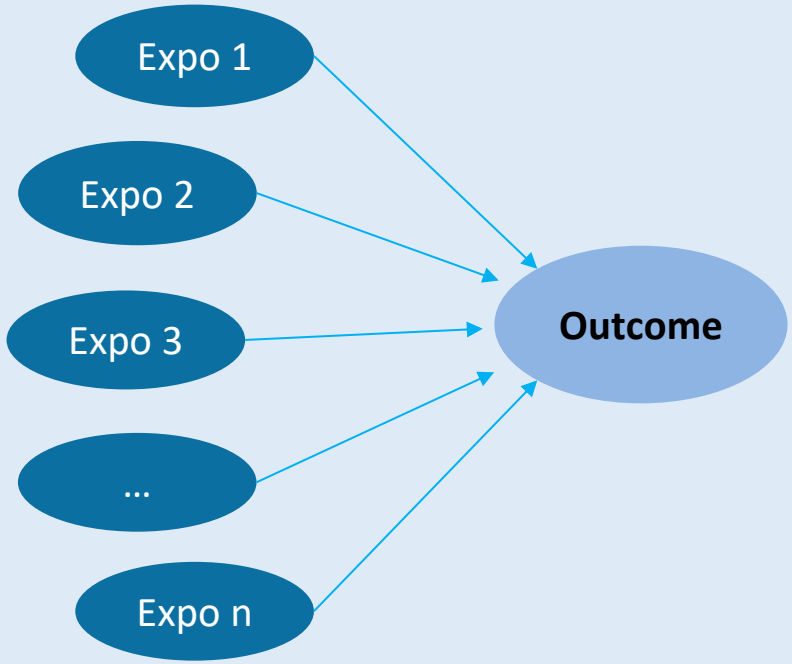
Association between exposures and health outcome

- In this section we are interested in studying how the **different exposures** are related to a health outcome
- We want to interpret the **effect of each factor individually**, but ideally considering the other exposures
- Regression models are usually adequate for this purpose
- But, in the context of the exposome, we will have to deal with:
  - Problems arising from **multiple testing**
  - **Correlation** between exposures
  - Potential **confounding** between exposures

1. ExWAS approach: generalities

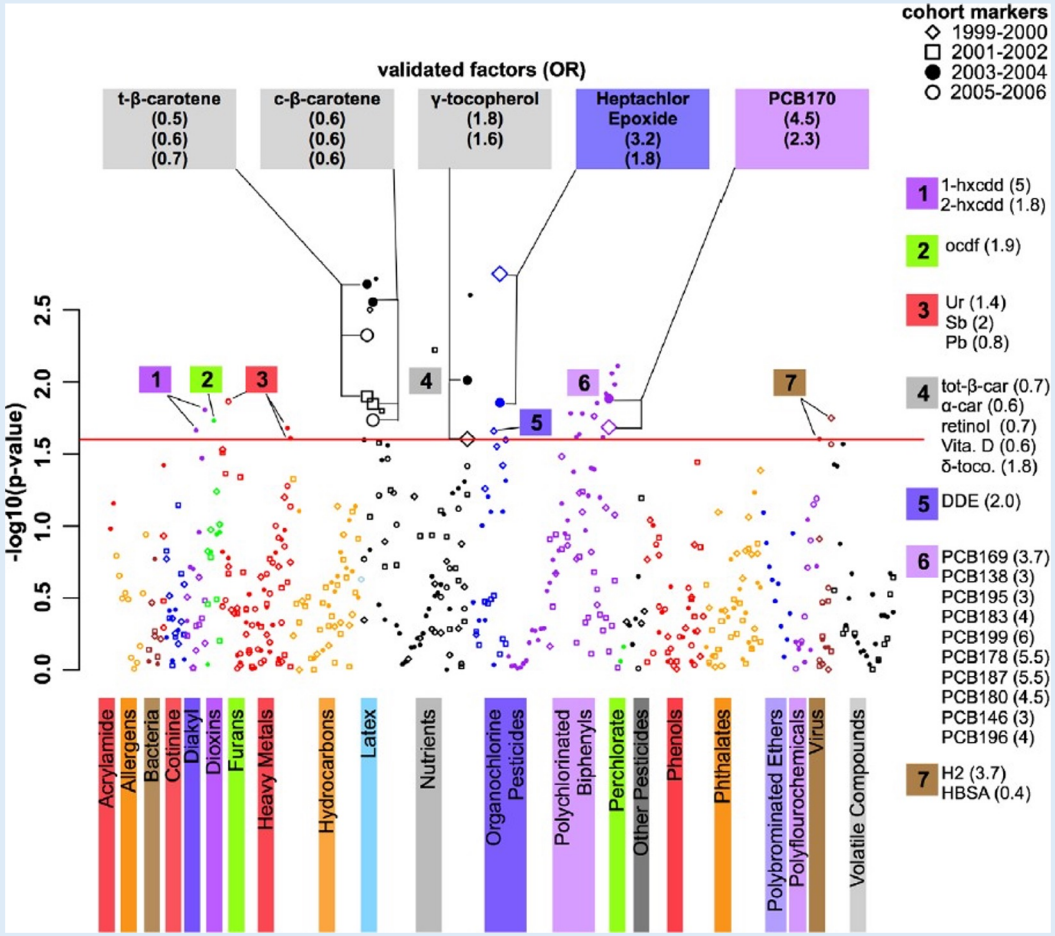
- **Exposure Wide Association Study** (ExWAS)
- Technique derivative from the **omics science**
- **How it works?**
  - Estimate the association between each exposure and the outcome, in independent models

    = there are as many regression models as there are exposures
- Co-exposure are not considered (no adjustment)
- Need to correct for **multiple testing**



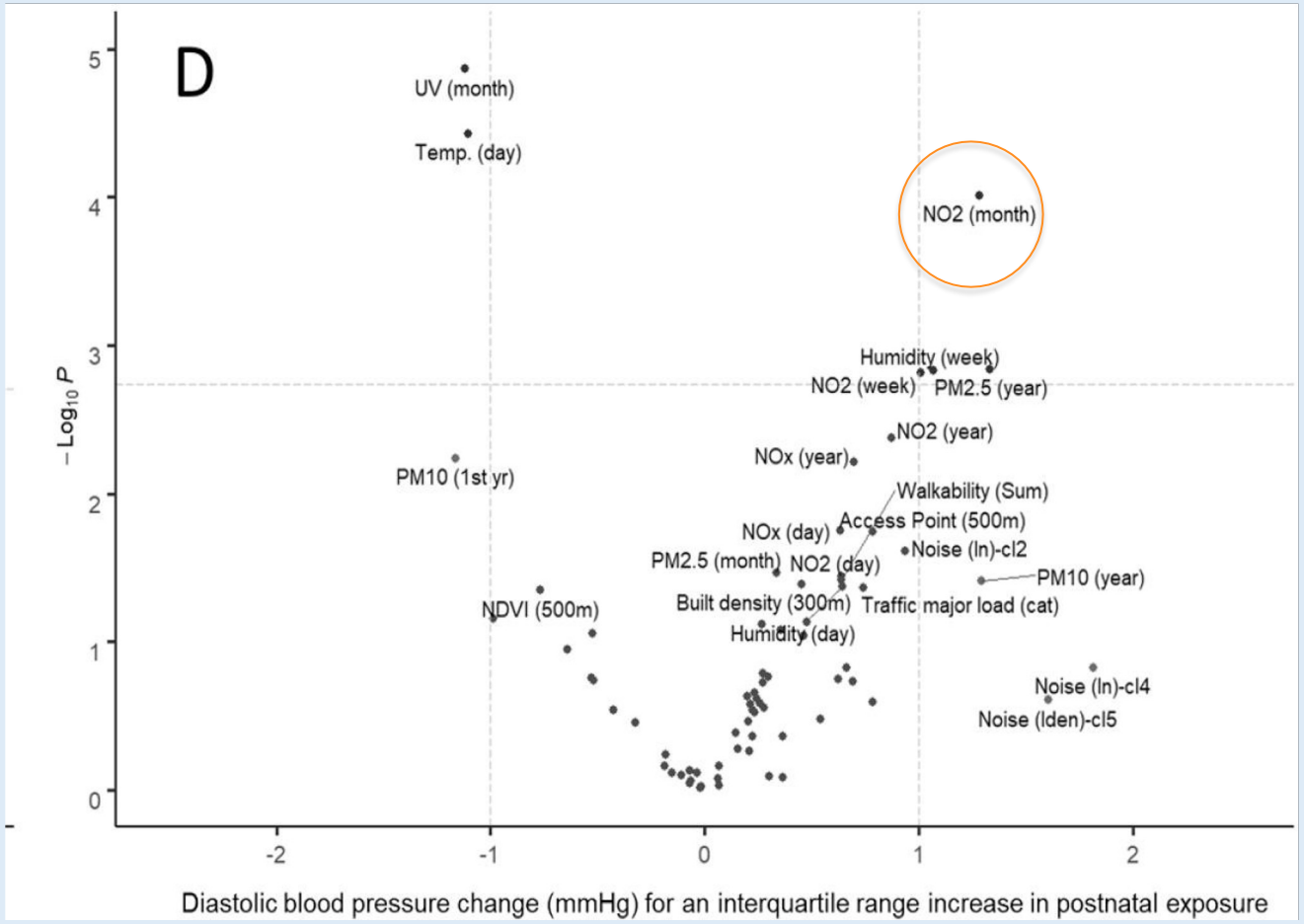
1. ExWAS approach: visualization of results

Manhattan plot



Patel, Plos One 2010

Volcano plot



Warembourg, Environ Int 2021

1. ExWAS approach: Pros & Cons

Pros

- High sensitivity
- Easy to implement (various regression models)
- Of utility in exploratory studies

Cons

- High FDR
- At risk of residual confounding
  - Same set of adjustment variables for all models
  - No adjustment for co-exposure (→ Multi-ExWAS?)

→ Need to be complemented by other methods

2. Variable selection methods

- Also called **subset regression** or **feature selection**.
- **Select a subset of relevant exposures** in a final model.
- Usually done automatically by algorithms (high-dimensionality).
- Associations are adjusted for exposures included in the final model.
- In principle, exposures not included in the final model are not associated with the outcome, thus they cannot be confounders. Still, selection is not that simple.
- Hundred of algorithms.

Harrell 2015

2. Variable selection methods: Stepwise selection

- Tries different models, includes or excludes variables as a function of p-values.

- Very popular, easy to use in statistical packages, but it has several problems and is **not recommended**:

  - it yields p-values that are too small.

  - estimated coefficients are biased towards inflated effects

  - final model prone to residual confounding

Harrell 2015

2. Variable selection methods: Penalized methods - LASSO

- **LASSO** method: the sum of the absolute value of the (scaled) coefficients in the final model is constrained to be less than some constant k chosen by cross-validation

- This forces some regression coefficients to be 0.

- The selected coefficients are **shrunk**, avoiding overfitting (optimism) in estimation.

2. Variable selection methods: Penalized methods - ENET

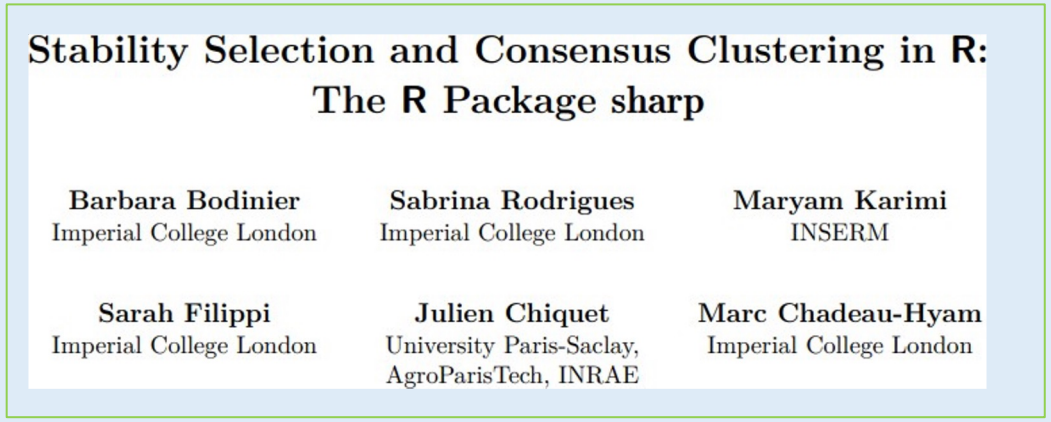
- With highly correlated variables, the LASSO tends to select one variable from a group and ignore the others.
- **Elastic net** is a method that combines the LASSO penalty with a Ridge penalty, which penalizes the sum of squared coefficients, thus penalizing large coefficient estimates.
- With highly correlated variables, Ridge tends to select several of them with similar magnitude.
- Elastic net seems a good compromise.

2. Variable selection methods: DSA algorithm

- Iterative algorithm similar to stepwise, but it uses **cross-validation** to decide the **deletion**, **substitution** or **addition** of a variable in the model.
- The selected variables need to be included in a classical regression model for estimate interpretation
- Subject to **instability**. It is recommended to run the DSA algorithm several times and select the exposures that were retained in several of them (e.g., 50%)

2. Variable selection methods: Stability of regularized methods

- Penalization parameters are chosen via **cross-validation**, which involves a random component.
- This may lead to problems of **stability** – even with the same data the method does not always returns the same selection of variables.
- There are stabilization algorithms, but they affect the performance of the method.



Stability Selection and Consensus Clustering in R: The R Package sharp

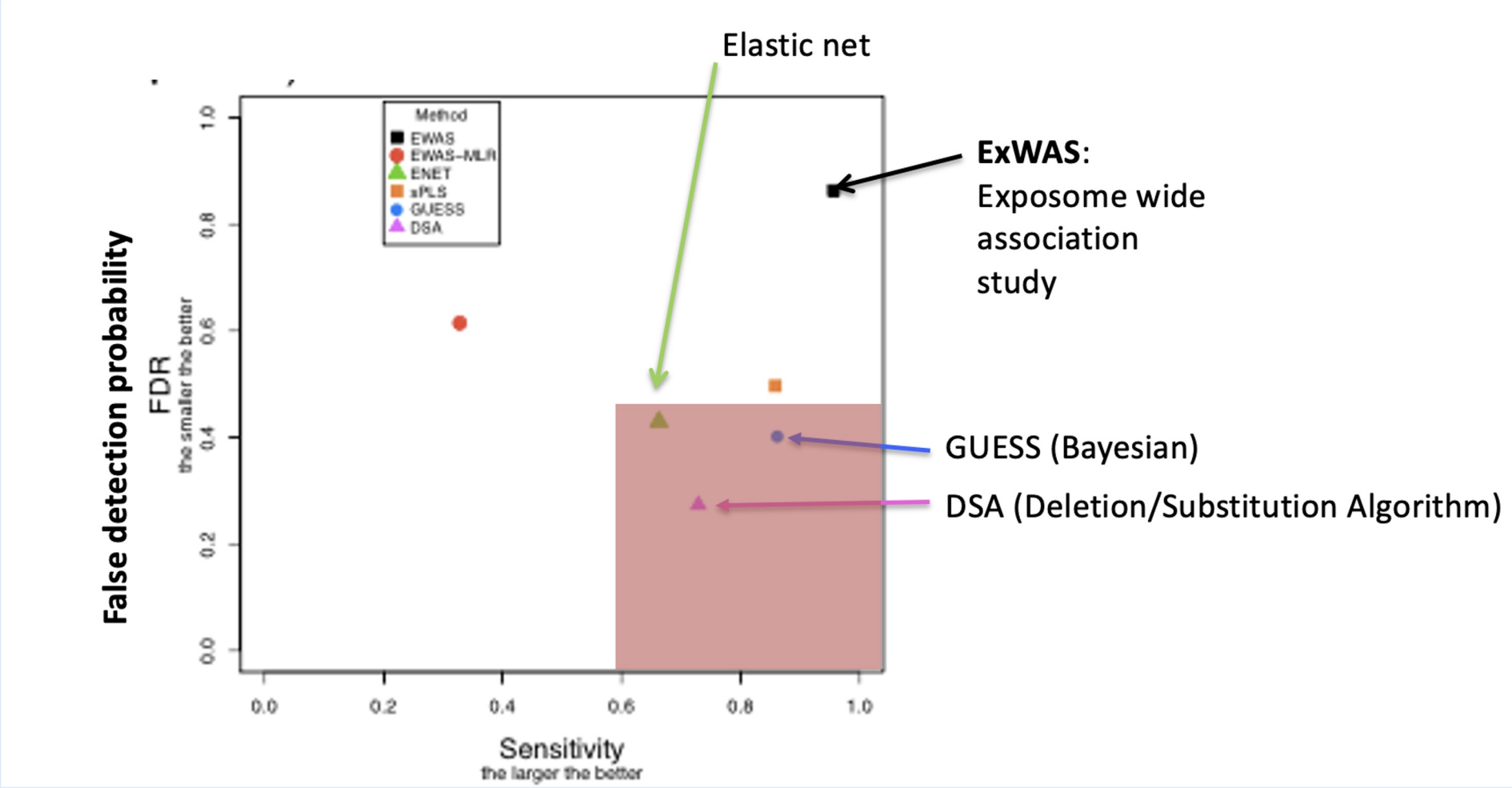
Barbara Bodinier, Imperial College London
Sabrina Rodrigues, Imperial College London
Maryam Karimi, INSERM
Sarah Filippi, Imperial College London
Julien Chiquet, University Paris-Saclay, AgroParisTech, INRAE
Marc Chadeau-Hyam, Imperial College London

2. Variable selection methods: comparison of methods FDR

- Simulation in some particular exposome setting (>200 vars.)

2. Variable selection methods: comparison of methods FDR

- Simulation in some particular exposome setting (>200 vars.)



Agier EHP 2016

2. Variable selection methods: extensions

- These algorithms also have extensions to search for
  - Interactions
  - Non-linear effects
  - Use biological or external information to aid in selection
  - Consider or take into account measurement error
  - Work with repeated measures

2.2 Association analysis

Are there specific exposure profiles associated with the outcome?

Cluster analysis

- The aim is to identify groups or clusters of subjects that **share a similar exposome** (exposures), and to evaluate whether these clusters are associated with the outcome

- Clustering techniques seek groups that
  - **minimize within-group variability** (clusters are homogeneous), and
  - **maximize between-group variability** (clusters are as different as possible from each other)

Cluster analysis

- Two different approaches:
  - **Unsupervised**: only data from exposures are used. The resulting clusters are then tested in association with the outcome
  - **Supervised**: the outcome is used to define the clusters. It identifies clusters (based on exposures) that, at the same time, show differences in the outcome
  - Different methods exist for cluster identification

Example of unsupervised clustering

- **Hierarchical clustering on principal components** (HCPC)

1. Principal component analysis (PCA)
   - Apply a PCA on your data
   - Identify the number of components needed to explain a certain amount of variance (e.g., 80%)
   - Run again a PCA, specifying the number of components you want to retain

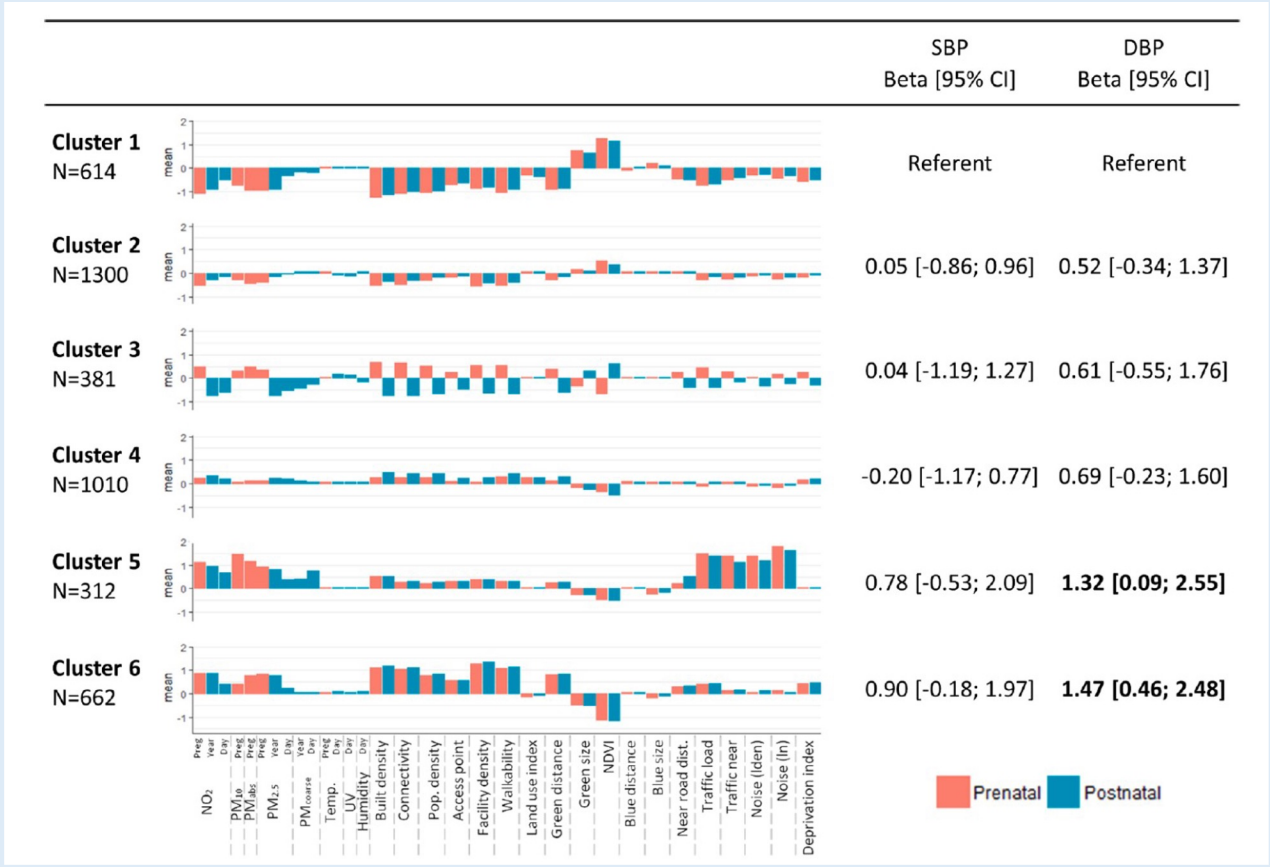
2. Hierarchical ascending classification on principal components
   - Apply a HCPC on the resulting components of the PCA to perform the cluster analysis
   - Identify the best number of clusters (graphically or automatically)
   - Interpret the clusters

3. Regression model
   - Include the cluster variable as an independent variable in a regression model

Example of unsupervised clustering

- **Hierarchical clustering on principal components** (HCPC)

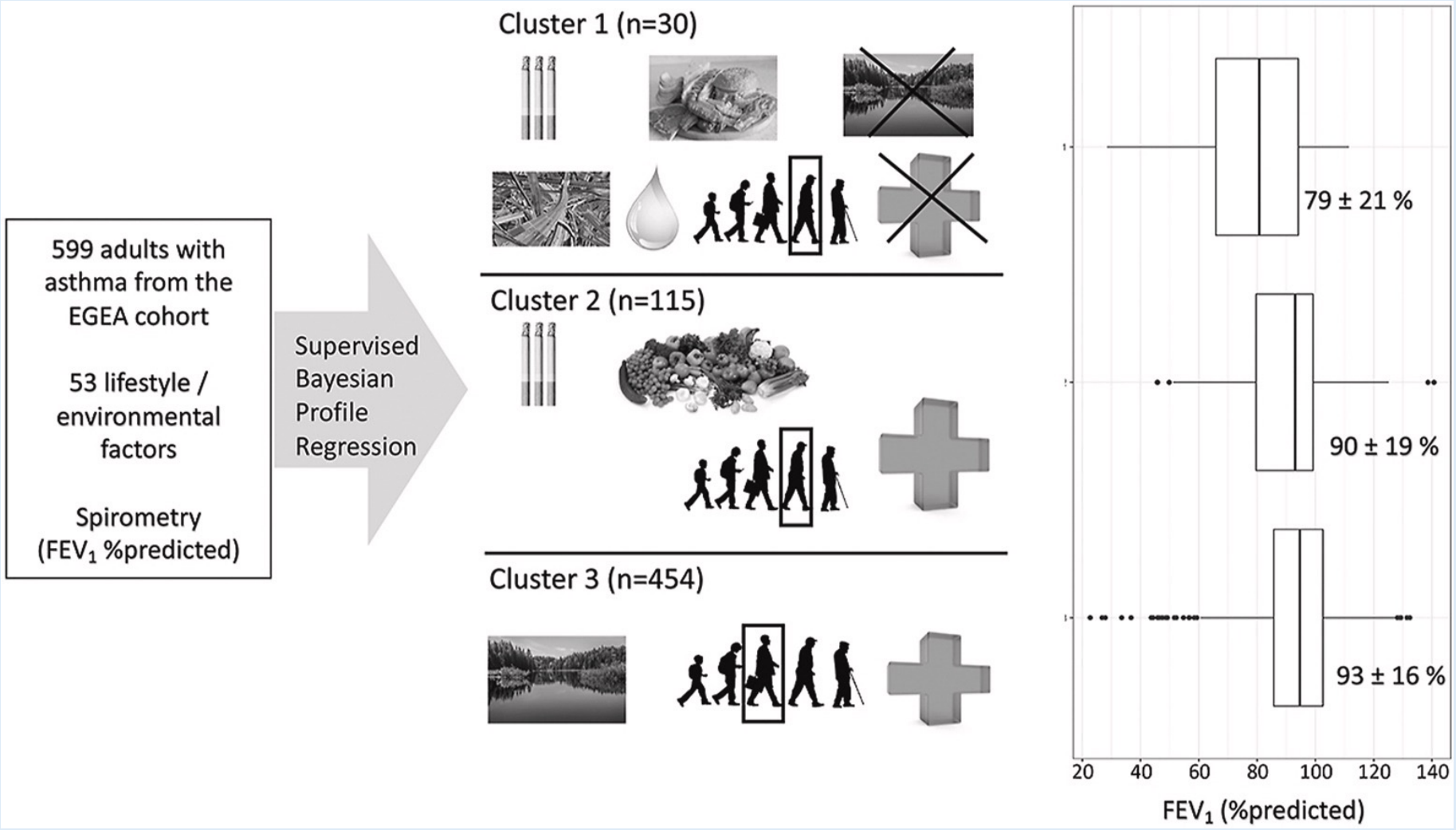
  - Helix study, n=5300
  - Urban exposome and child blood pressure
  - The clusters 5 and 6 = Higher exposure to air pollution and noise, less access to green spaces
  - Both associated with higher blood pressure



Warembourg, Environ Int 2021

Example of supervised clustering

- **Bayesian profile regression** – Example for lung function



Guillien Environ Res 2021

Clustering analysis: Pros & Cons

Pros

- All exposures are considered simultaneously
- Takes into account correlations between exposures
- Limits the number of tests
- Possibility of supervised and unsupervised methods

Cons

- Methods sometimes unstable
- Interpretation, small cluster size
- Confounding not easy to control

2.3

Association analysis

Are there mixtures of exposures that affect the outcome?

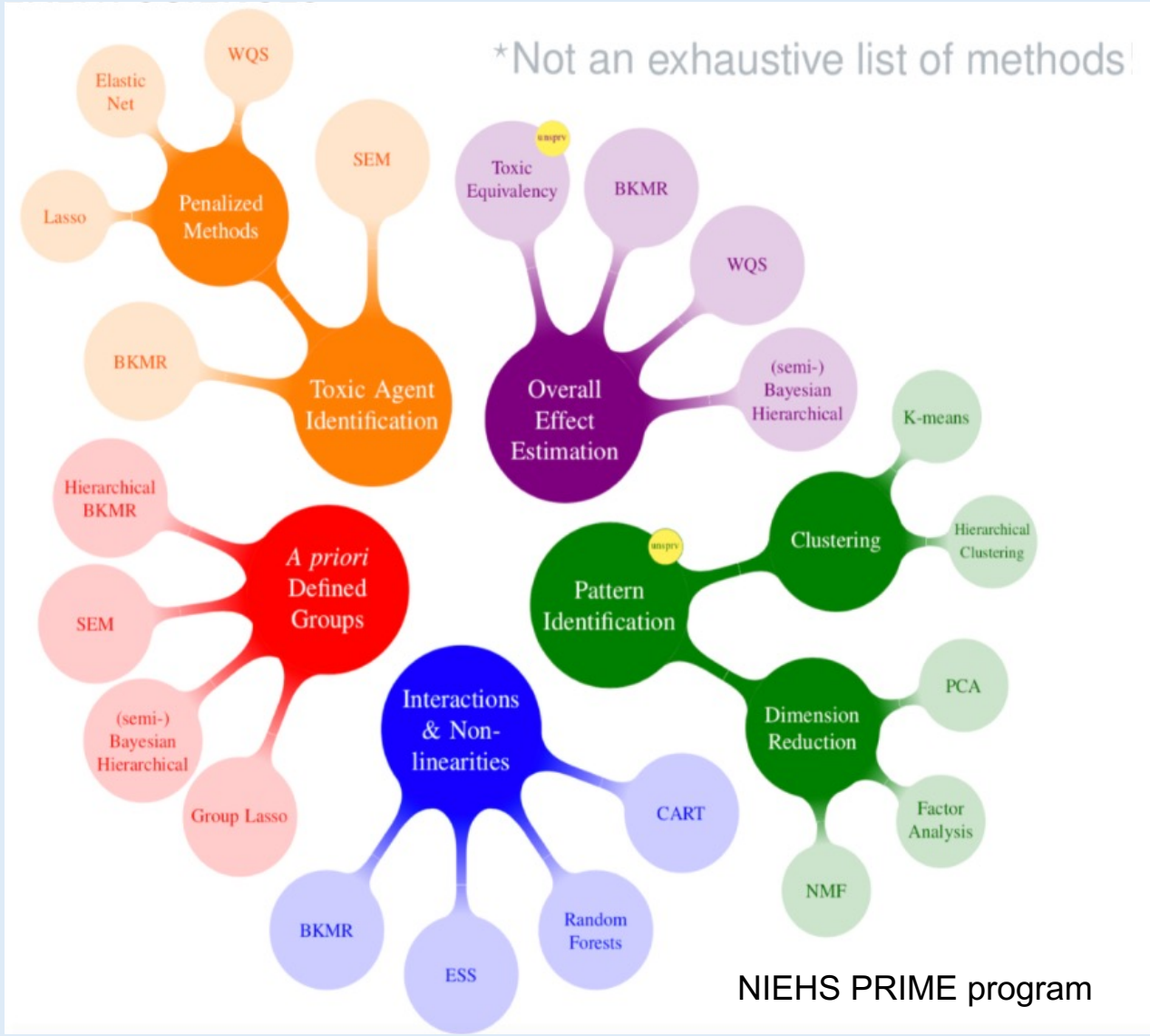
Statistical methods to explore mixture effect

The choice depends on the research question:

-Overall effect estimation?
-Identification of the most contributing exposures?
-Identification of interactions?
-Identification of exposure profile?
-Based on a priori knowledge ?
-...

Quantile g-computation (Qgcomp)
Weigthed Quantile Sum regression (WQS)
Bayesian Kernel Machine regression (BKMR)

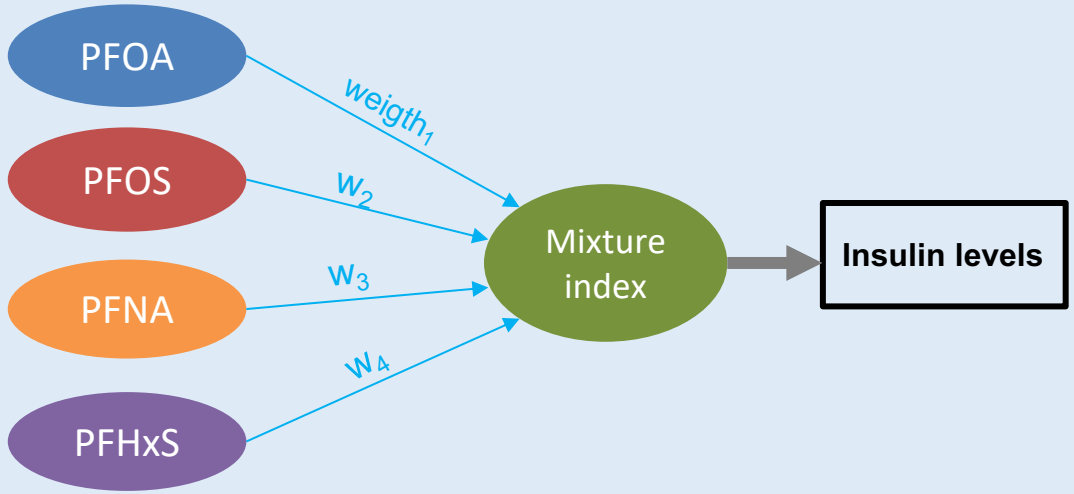
(Gibson E et al, Environmental Health 2019)
(Stafoggia M et al, Curr Environ Health Rep 2017)
(Hamra G et al, Curr Epidemiol Rep 2018)
(Keil A et al, Environ Health Perspect 2020



Weighted Quantile Sum Regression (WQS)

Main idea: To construct a weighted index estimating the mixed effect of all predictor variables on an outcome.

- Converts exposures into quantiles (to standardised the variables and avoid the effect of outliers)
- Calculate the sum of weights of all quartiles
- The weights add to 1 and are restricted to a single direction of effect on the outcome (all increasing or decreasing the outcome effect)
- The contribution of each individual predictor to the overall index effect can then be assessed by the relative strength of the weights the model assigns to each variable.



WQS: the basic function

```
library(gWQS)

mixture<-c("pfoa_m","pfos_m","pfna_m","pfhxs_m")

results1 <- gwqs(insulin_9yr_log2 ~ wqs + covariates,
                 mix_name = mixture,
                 data = data.expo,
                 validation = 0.6,
                 b = 100,
                 q = 4,
                 family="gaussian")
```

Vector with exposure names

Formula

Name of the mixture

Name of the dataset with original data

% of validation vs training data

Number of bootstrap sample

Categorization of exposure

Regression family

4 = quartiles

40% training vs 60% validation

validation=0 → entire dataset used to train and validate the model

WQS: the basic function

```
library(gWQS)

mixture<-c("pfoa_m","pfos_m","pfna_m","pfhxs_m")

results1 <- gwqs(insulin_9yr_log2 ~ wqs + covariates,
                 mix_name = mixture,
                 data = data.expo,
                 validation = 0.6,
                 b = 100,
                 q = 4,
                 family="gaussian",
                 b1_pos=TRUE, b1_constr=TRUE
                 seed=1024)
```

Vector with exposure names

Formula

Name of the mixture

Name of the dataset with original data

% of validation vs training data

Number of bootstrap sample

Categorization of exposure

Regression family

Force the association to be positive

Set seed to obtain reproducible results

WQS: mixture index

```
summary(results1$fit)
```

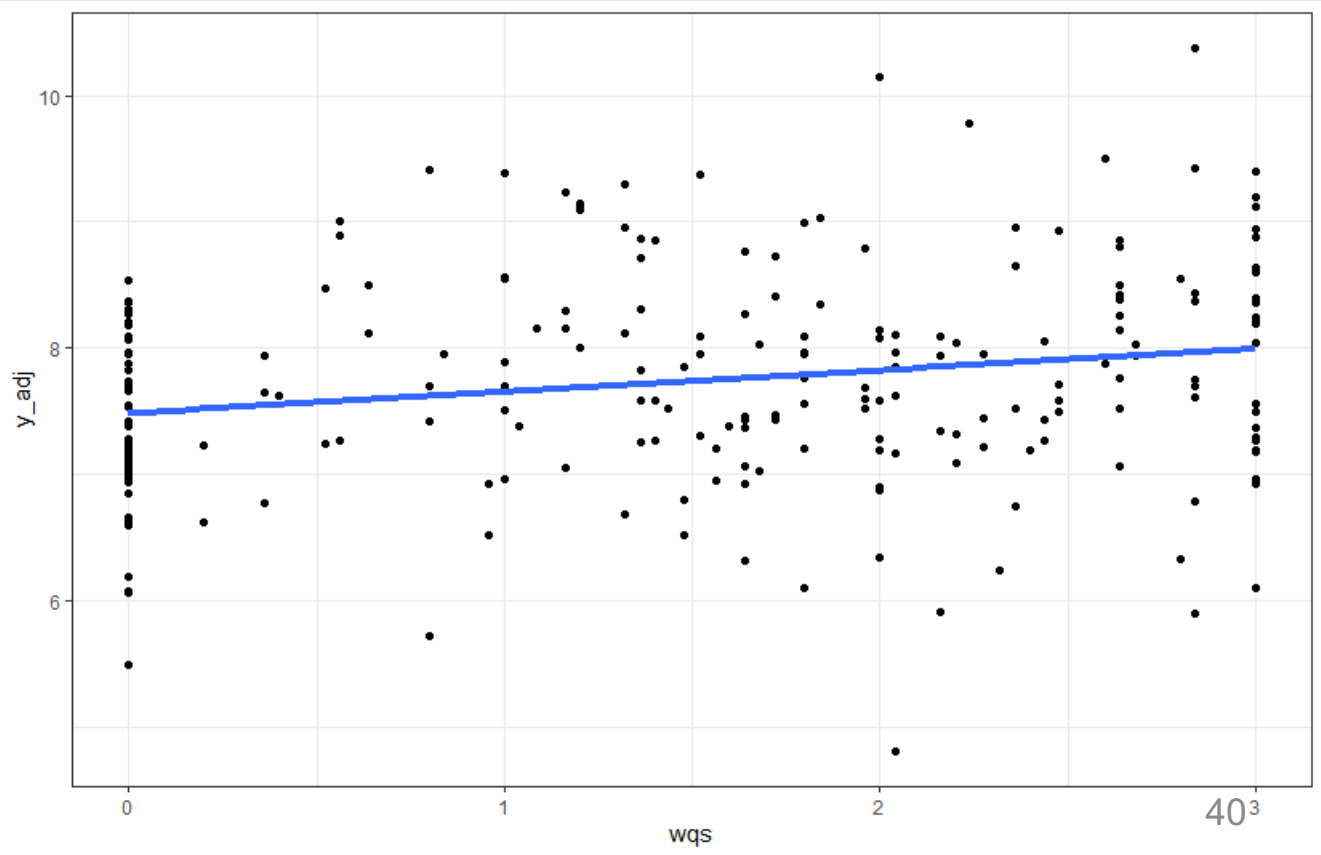
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.9014924  1.1409762  10.431  < 2e-16 ***
sex           -0.0674359  0.1201279  -0.561  0.57515
hs_c_age_9y   -0.4172316  0.1014437  -4.113 5.61e-05 ***
estudios3c_imp2 -0.1105208  0.1497322  -0.738  0.46126
estudios3c_imp3 -0.0562391  0.1669388  -0.337  0.73654
imcm           0.0001136  0.0123659   0.009  0.99268
paridad3c_imp2  0.1556759  0.1429443   1.089  0.27737
paridad3c_imp3  0.3193153  0.3371312   0.947  0.34465
edadm_imp      -0.0190784  0.0176316  -1.082  0.28047
m_not_eur_imp2  0.1453840  0.2792698   0.521  0.60320
wqs            0.2027949  0.0625108   3.244  0.00137 **
```

```
confint(results1$fit)
```

```
                    2.5 %      97.5 %
(Intercept)      9.66522014 14.13776462
sex             -0.30288233  0.16801056
hs_c_age_9y     -0.61605764 -0.21840564
estudios3c_imp2 -0.40399052  0.18294894
estudios3c_imp3 -0.38343322  0.27095496
imcm            -0.02412316  0.02435042
paridad3c_imp2  -0.12448972  0.43584151
paridad3c_imp3  -0.34144981  0.98008037
edadm_imp       -0.05363581  0.01547897
m_not_eur_imp2  -0.40197477  0.69274271
wqs              0.08027588  0.32531389
```

An quartile increase in the mixture index is associated with a 0.20 [0.08;0.33] increase in Insulin levels

```
gwqs_scatterplot(results1)
```

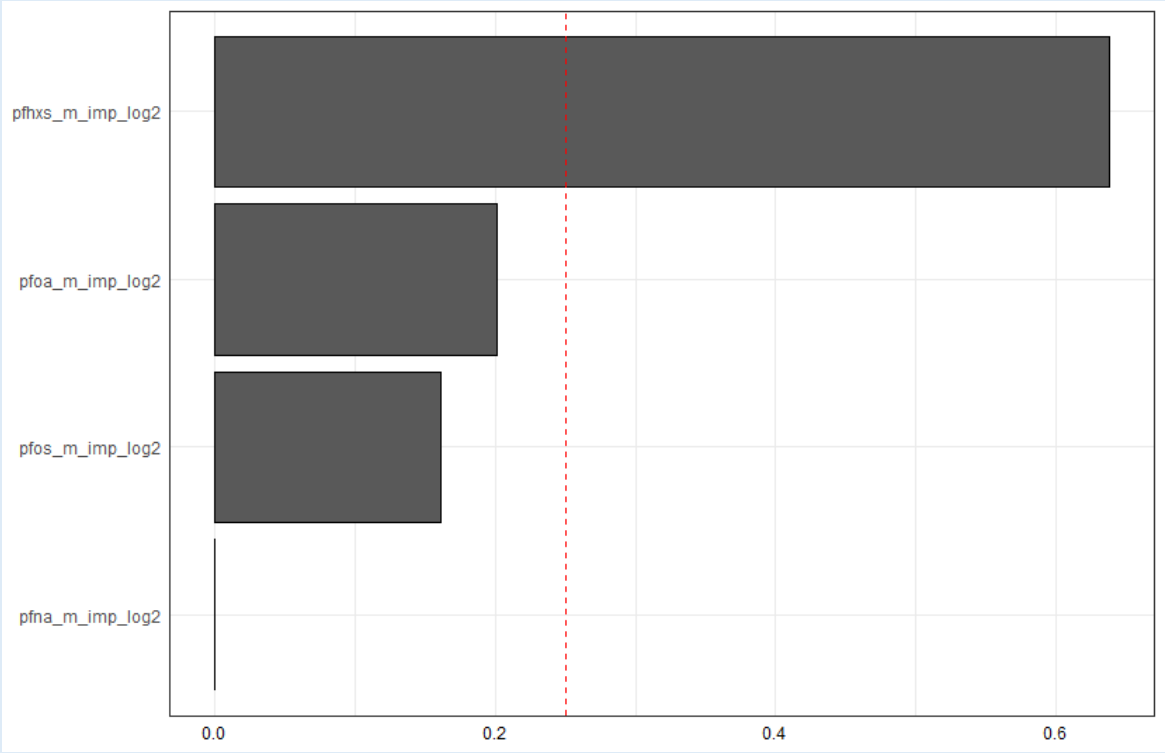


WQS: contributors to the mixture effect

```
results1$final_weights
```

	mix_name	mean_weight
pfhxs_m_imp_log2	pfhxs_m_imp_log2	0.638467954792251
pfoa_m_imp_log2	pfoa_m_imp_log2	0.200779801449394
pfos_m_imp_log2	pfos_m_imp_log2	0.160752242130693
pfna_m_imp_log2	pfna_m_imp_log2	0.000000001627661

```
gwqs_barplot(results1)
```



Threshold to discriminate which element has a 'significant' weight greater than zero:

τ = 100/number of chemicals

= 100/4

= 0.25

PFHxS is the largest contributor (64%) to the mixture effect

/!\ Result stability

- Two steps depend on random number (=seed)
  - The cross-validation
  - The weight of the estimate are estimated by bootstrap
- Repeating X times the model would lead to different results (except if we fix the seed but…)

Seed = 1024

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.2607     1.5779  38.824   <2e-16 ***
wqs          -0.6955     0.9416  -0.739    0.462
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```

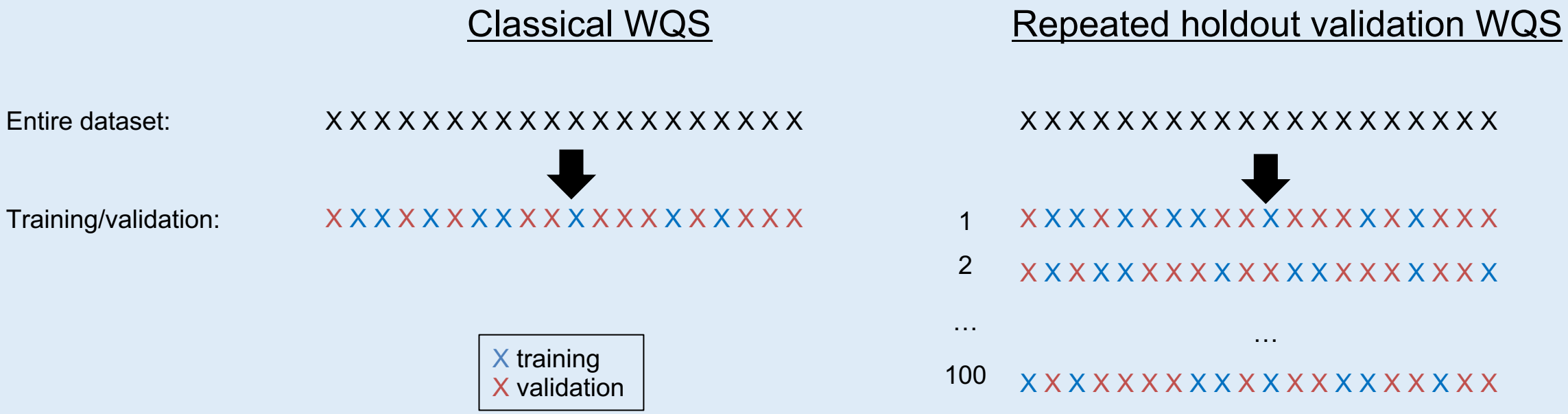
Seed = 99

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.9309     1.2185  50.826   <2e-16 ***
wqs          -1.5077     0.6967  -2.164   0.0323 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0
```

→ Unstable results

Repeated holdout validation for WQS

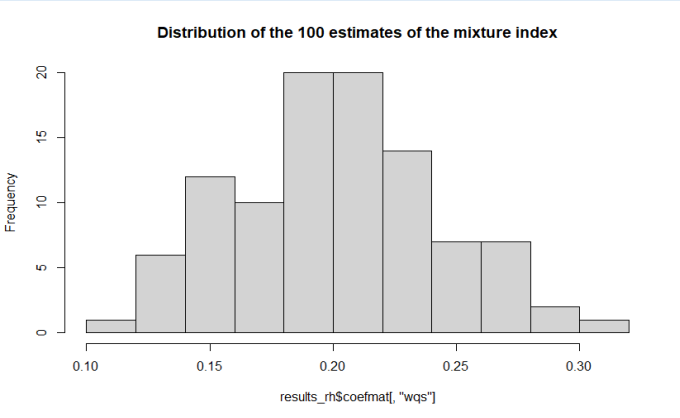
- General principal: This bootstraps the single-partition WQS 100 times to create a distribution of validated results



Repeated holdout validation for WQS

```
results_rh = gwqsrh(insulin_9yr_log2 ~ wqs + covariates, mix_name = mixture,
data = data.expo,validation = 0.6,q = 4, b=100, rh=100, seed=1024)
```

```
hist(results_rh$coefmat[,"wqs"])
```



If not approximaly normal, increase the number of iterations

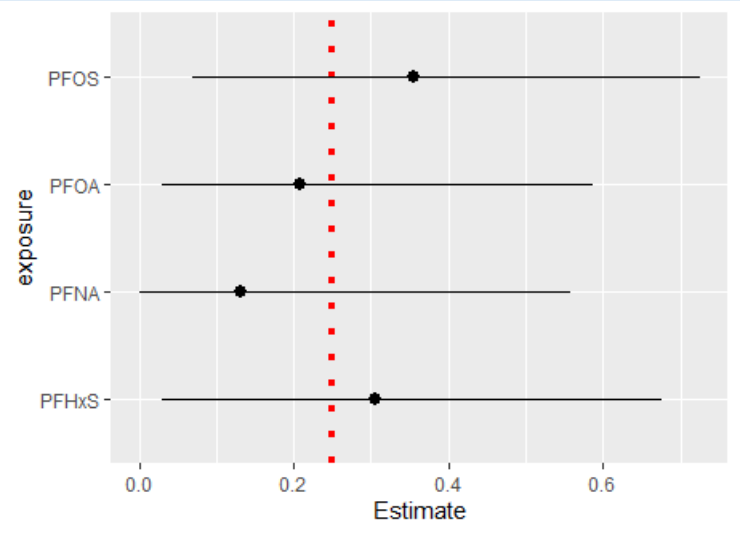
```
summary(results_rh)
```

```
Coefficients:
                  Estimate  Std. Error       2.5 % 97.5 %
(Intercept)     12.16011760 0.60208165 10.98003757 13.340
sex             -0.06561648 0.08501582 -0.23224748  0.101
hs_c_age_9y     -0.45099641 0.06184971 -0.57222184 -0.330
estudios3c_imp2 -0.04891917 0.10243776 -0.24969718  0.152
estudios3c_imp3 -0.08408266 0.11042103 -0.30050787  0.132
imcm            -0.00003744 0.00886206 -0.01740708  0.017
paridad2c_imp2   0.08631292 0.09036809 -0.09080854  0.263
edadm_imp       -0.01577519 0.00868720 -0.03280210  0.001
wqs              0.20234355 0.04068870  0.12259369  0.282
```

A quartile increase in the mixture index is associated with a 0.20 [0.12;0.28] increase in Insulin levels

```
results_rh$final_weights
```

```
                Estimate   2.5 %   97.5%
pfos_m_imp_log2  0.35473 0.06899 0.72477
pfhxs_m_imp_log2 0.30613 0.02976 0.67657
pfoa_m_imp_log2  0.20827 0.02920 0.58745
pfna_m_imp_log2  0.13087 0.00000 0.55725
```



PFOS and PFHxS are the main contributors

WQS: pros and cons

Pros

- Can deal with high number of exposures (>10)
- Easy to implement and interpret
- Available for various type of outcomes
- Reduce the impact of outliers (in exposure levels)

Cons

- Constraint in the direction of the effects of each individual exposure (either + or -)
- Loss of information (exposure categorization)
- Interaction are not supported
- Missing data are not supported

WQS: extensions

- **Bayesian WQS** (Colicino E et al, Environ Epidemiol 2020)
  - Infer the estimates using the whole dataset
  - No assumption on directional homogeneity

- **Grouped WQS** (Wheeler D et al, EHP 2016; groupWQS vignette)
  - Allow to place chemicals into groups
  - Different direction of the effect can be determined for each pre-defined group of chemicals

Quantile g-computation (qgcomp)

Main idea: To estimate the effect of a mixture of exposures on an outcome

Categorise exposures (usually quartiles)

↓

Sum of the adjusted coefficients

↓

$$Y_i = \beta_0 + \psi \sum_{j=1}^{p} w_j X_j^q + \varepsilon_i$$

The effect on the outcome of increasing every exposure by one quantile

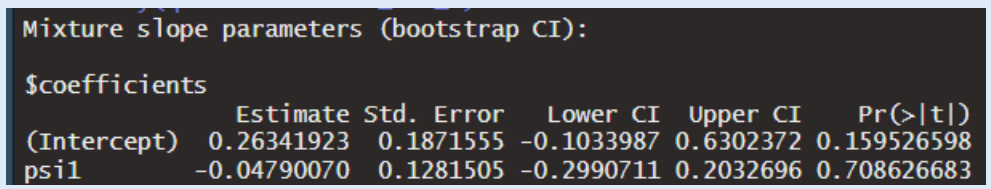
When all effects are linear and in the same direction ("directional homogeneity"), quantile g-computation is equivalent to weighted quantile sum regression in large samples.

Qgcomp: the basic function

```
metal_gcomp <- qgcomp(Groupe ~ Pb + Hg + Cr + Cd +
                                Age + Centre + Tabac +  Education_classe +
                                Saison + Origine_classe, # complete formula including exposures
                      expnms = c("Pb", "Hg", "Cr", "Cd"), #list of exposures
                      data = imp_data,
                      family = binomial(),
                      rr = FALSE) #if using binary outcome and rr=TRUE (default), estimates
risk ratio rather than odds ratio
```

```
summary(qgcomp_fit)
```

(not the actual example)



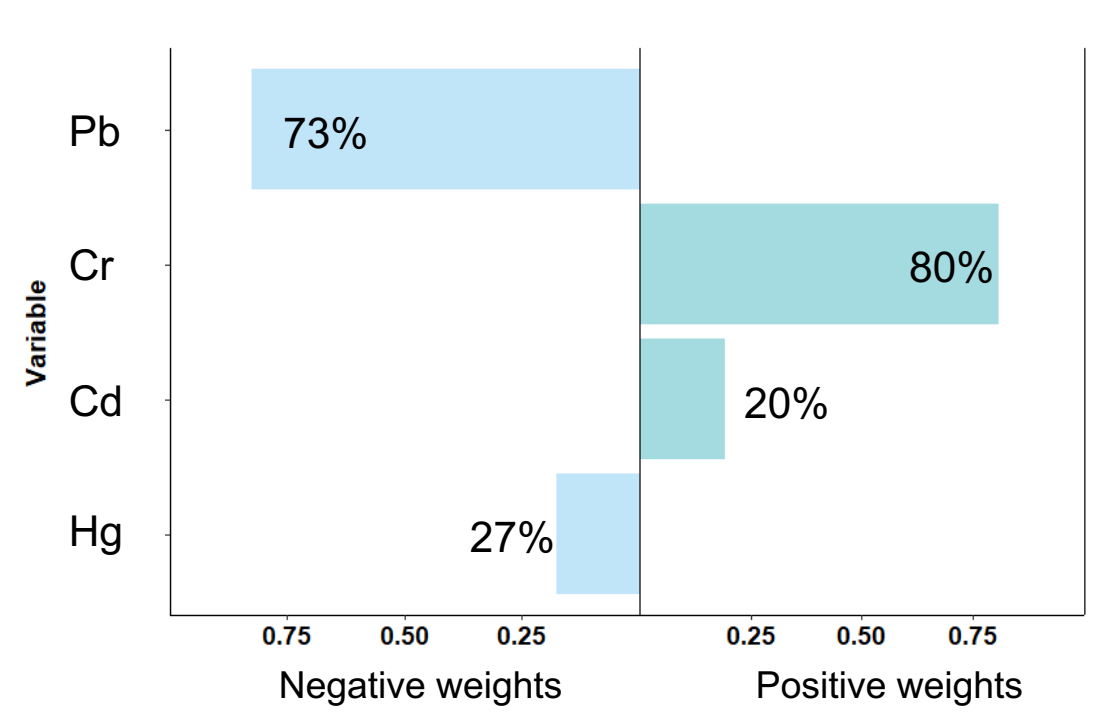
```
Mixture slope parameters (bootstrap CI):

$coefficients
             Estimate Std. Error   Lower CI  Upper CI    Pr(>|t|)
(Intercept)  0.26341923  0.1871555 -0.1033987 0.6302372 0.159526598
psi1        -0.04790070  0.1281505 -0.2990711 0.2032696 0.708626683
```

Overall Effect

OR = 1.40, 95% CI [0.85 - 2.32]

Qgcomp: the basic function

plot(qgcomp_fit) #weights



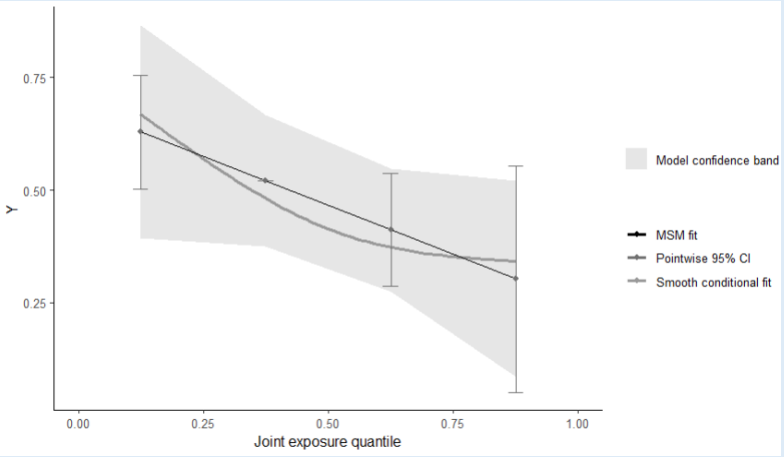
The weights are the sum of the coefficients divided by the total mixture effect.

Qgcomp: non-linearity

Non linearity of single exposure effects

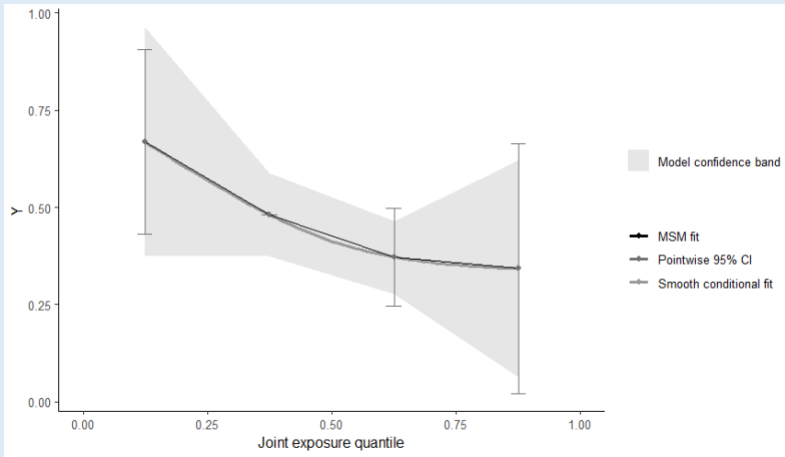
```
list_metal<-c("hs_cd_c_Log2","hs_hg_c_Log2","hs_pb_c_Log2")

qcboot.fit4 <- qgcomp(hs_zbmi_who~. + .^2,
                      expnms=list_metal,
                      data = data[,c("hs_zbmi_who","h_cohort",
                      "e3_sex_None","e3_yearbir_None", list_metal)],
                      family=gaussian(),
                      q=4,
                      seed=125,
                      rr = FALSE)
```

Only variables in the formula



Non linearity of mixture effect

```
list_metal<-c("hs_cd_c_Log2","hs_hg_c_Log2","hs_pb_c_Log2")

qcboot.fit5 <- qgcomp(hs_zbmi_who~. + .^2,
                      expnms=list_metal,
                      data = data[,c("hs_zbmi_who","h_cohort",
                      "e3_sex_None","e3_yearbir_None", list_metal)],
                      family=gaussian(),
                      degree=2,
                      q=4,
                      seed=125,
                      rr = FALSE)
```



Qgcomp: pros and cons

Pros

- Computational efficiency
- No directional homogeneity required
- Non-linearity possible

Cons

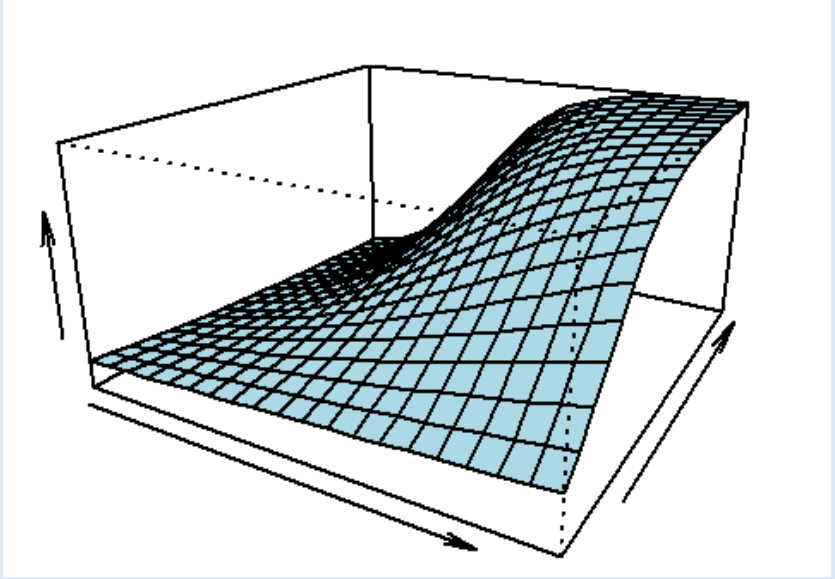
- Loss of information (exposure categorization)
- Multiple imputation not integrated
- Interactions possible on a small number of exposures

Bayesian Kernel Machine Regression (BKMR)

- Main idea: to model the health outcome as a smooth function *h* of the exposure variables, adjusted for possible confounders

h is called exposure-response function, and is estimated with a Kernel function

Variable selection can be performed (or not)



$$y_i = h(x_{i1}, \ldots, x_{iP}) + \mathbf{z_i}^T \gamma + \epsilon_i,$$

BKMR: preparations

- Data pre-processing
  - Remove (or substitute) any missing data in covariate, exposure or outcome variables
  - Center/scale continuous variables and create dummy variables for categorical variables

```
data.expo$estudio01   <- as.numeric(data.expo$estudios3c_imp==2)
data.expo$estudio02   <- as.numeric(data.expo$estudios3c_imp==3)
data.expo$pfoa_m_imp_log2_z   <- scale(data.expo$pfoa_m_imp_log2)
 …
```

  - Create 3 object/matrix:

    The outcome (Y)

```
outcome<-data.red$insulin_9yr_log2
```

    The exposures (Z)

```
mixture<-with(data.red,cbind(pfoa_m_imp_log2_z,pfos_m_imp_log2_z,
                             pfna_m_imp_log2_z,pfhxs_m_imp_log2_z))
colnames(mixture)<-c("PFOA","PFOS","PFNA","PFHxS")
```

    The covariates (X)

```
covariates <- with(data.red, cbind(sex01, estudio01, estudio02, paridad01,
hs_c_age_9y_z, imcm_z, edadm_imp_z))
```

BKMR: the basic function

```
library (bkmr)
fit <- kmbayes(y=outcome,
               Z=mixture,
               X=covariates,
               iter=20000,
               verbose=TRUE,
               varsel=TRUE,
               family="gaussian",
               id=NULL,
               ztest=NULL)
```

Define the outcome, the exposures and the covariates (see previous slide)

Number of iterations

Display diagnostic information during the process

Indicate if variable selection is perfomed (TRUE)

Type of outcome

Vector indicating a group variable in the case of hierarchical data

Vector indicating on which variables Z to conduct variable selection

BKMR: PIPs

- The **posterior inclusion probability** is a ranking measure (0 to 1) to see how much the data favors the **inclusion** of a variable in the regression.

ExtractPIPs(fit)

variable	PIP
PFOA	0.1080
PFOS	0.7000
PFNA	0.1812
PFHxS	0.1744

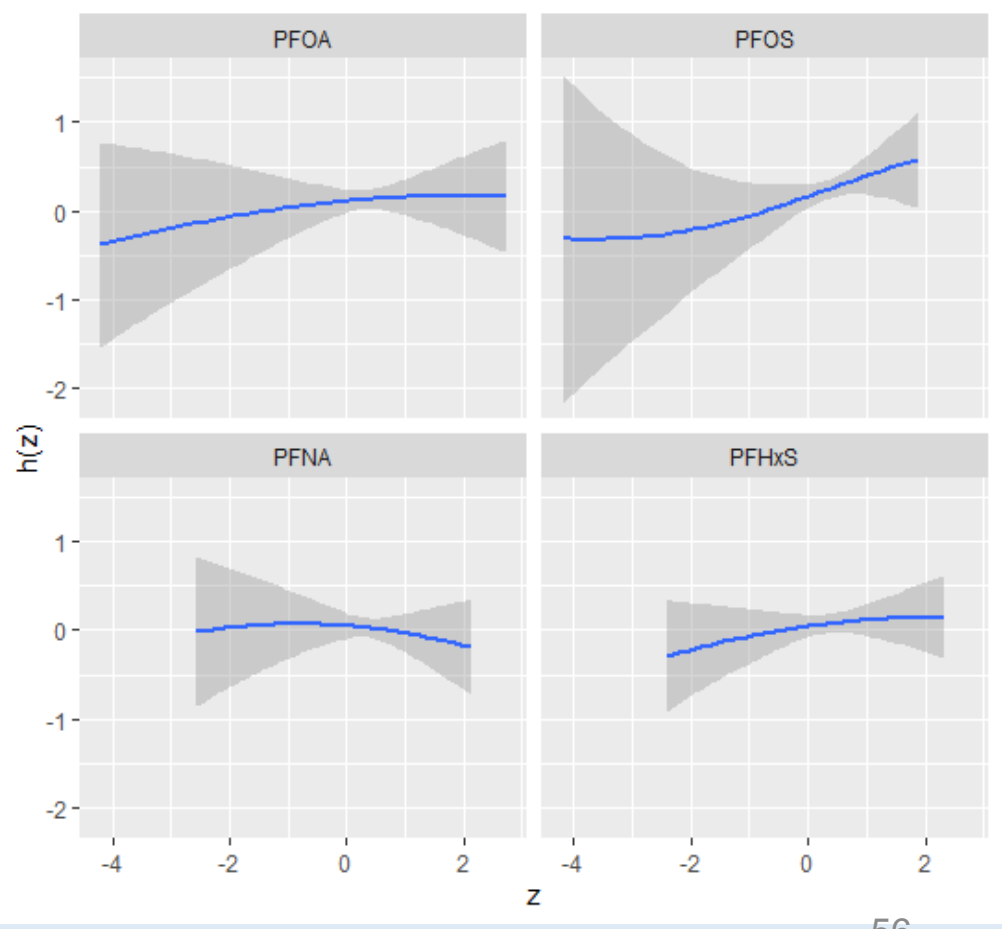
All 4 PFAS were selected (none shrinked to 0)

PFOS is likely to be a true predictor (70%)

BKMR: single-exposure dose-response relationships

- To visualize the relationship between 1 exposure and the outcome while fixing the other to a specific value (e.g. the median)

```
pred.resp.univar <- PredictorResponseUnivar(fit = fit)

ggplot(pred.resp.univar, aes(z, est, ymin = est-1.96*se,
  ymax = est + 1.96*se)) +
  geom_smooth(stat = "identity") +
  facet_wrap(~ variable) +
  ylab("h(z)")
```

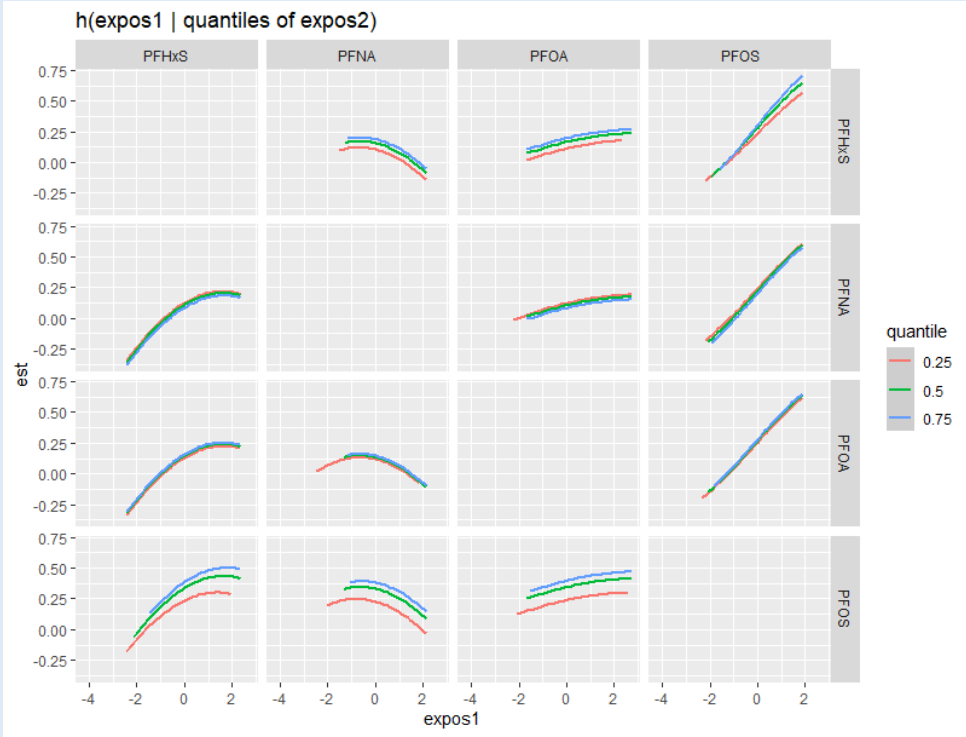


- Note that BKMR can capture non-linear associations

BKMR: bivariate dose-response relationships

- To investigate the exposure-response function of a single exposure where the second exposure is fixed at various quantiles

```
pred.resp.bivar.levels <- PredictorResponseBivarLevels(
  pred.resp.bivar, mixture, qs = c(0.25, 0.5, 0.75))

ggplot(pred.resp.bivar.levels, aes(z1, est)) +
  geom_smooth(aes(col = quantile), stat = "identity") +
  facet_grid(variable2 ~ variable1) +
  ggtitle("h(expos1 | quantiles of expos2)") +
  xlab("expos1")
```



- Interpretation of the plot (examples):
  - PFOS shows the strongest effect
  - No evidence of interaction (parallel curves)

BKMR: Summary statistics of the overall predictor-response function

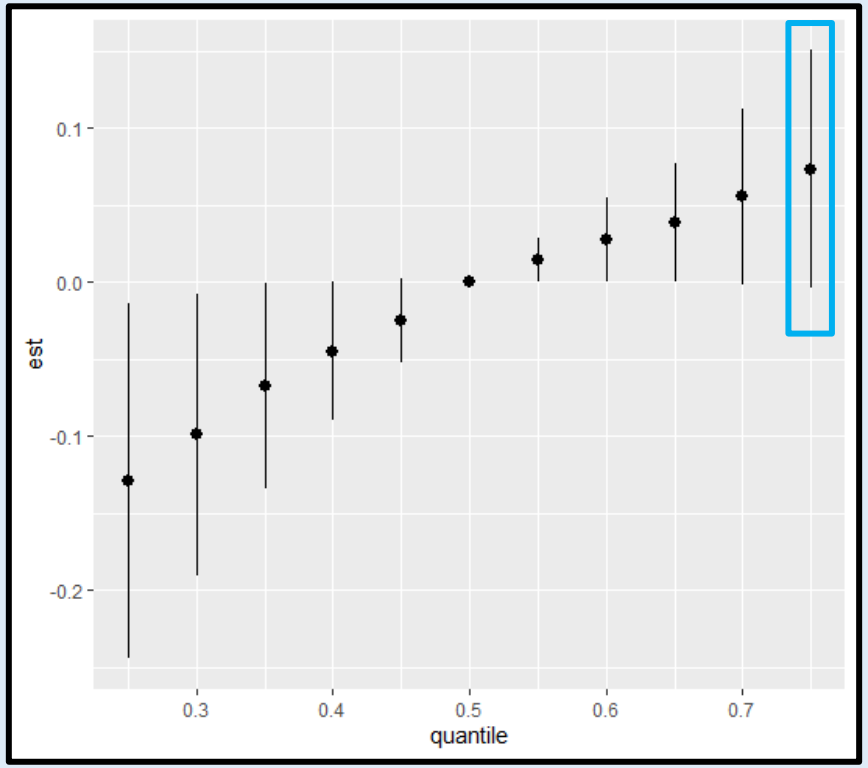
- To estimate the overall effect of the mixture when all exposures are at a particular percentile as compared to when all of them are at their 50th percentile

```
risks.overall <- OverallRiskSummaries(fit = fit, y =
outcome, Z = mixture, X = covariates, qs = seq(0.25,
0.75, by = 0.05), q.fixed = 0.5, method = "exact")
```

```
risks.overall
```

quantile	est	sd
0.25	-0.12913598	0.058883483
0.30	-0.09910316	0.046659379
0.35	-0.06769920	0.033862374
0.40	-0.04500291	0.022911622
0.45	-0.02519523	0.013756598
0.50	0.00000000	0.000000000
0.55	0.01401333	0.007198663
0.60	0.02732083	0.013749458
0.65	0.03851379	0.019760012
0.70	0.05521417	0.028999797
0.75	0.07298084	0.039375659

```
ggplot(risks.overall, aes(quantile, est,
ymin = est - 1.96*sd, ymax = est + 1.96*sd)) +
geom_pointrange()
```



Overall mixture effect = 0.07 [-0.004;0.15] when all exposures are fixed at the 75th percentile as compared to when all of them are fixed at the 50th percentile

BKMR: Summary statistics of the univar predictor-response function

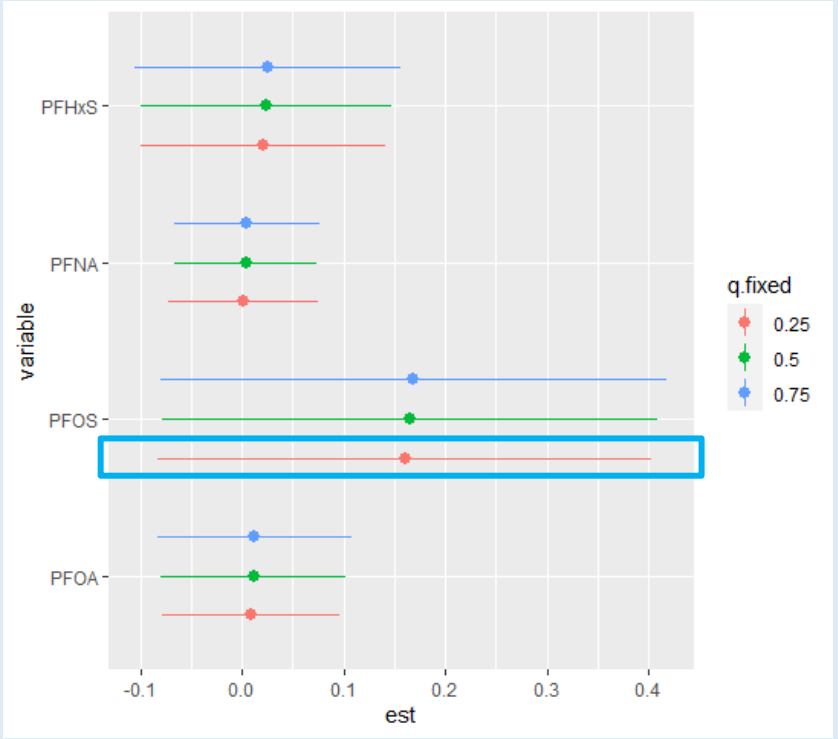
- To estimate the individual effect of a specific exposure (e.g., comparing the 75th to the 25th percentile) when all the others are fixed at a particular percentile (e.g., 25, 50 and 75th)

```
risks.singvar <- SingVarRiskSummaries(fit = fit, y =
outcome, Z = mixture, X = covariates, qs.diff = c(0.25,
0.75), q.fixed = c(0.25,0.50,0.75), method = "exact")
```

```
risks.singvar
```

q.fixed	variable	est	sd
<fct>	<fct>	<dbl>	<dbl>
0.25	PFOA	0.00895	0.0445
0.25	PFOS	0.160	0.124
0.25	PFNA	0.00189	0.0375
0.25	PFHxS	0.0208	0.0611
0.5	PFOA	0.0113	0.0464
0.5	PFOS	0.165	0.125
0.5	PFNA	0.00378	0.0359
0.5	PFHxS	0.0240	0.0631
0.75	PFOA	0.0127	0.0490
0.75	PFOS	0.169	0.127
0.75	PFNA	0.00484	0.0365
0.75	PFHxS	0.0259	0.0666

```
ggplot(risks.singvar, aes(variable, est, ymin = est -
1.96*sd, ymax = est + 1.96*sd, col = q.fixed)) +
geom_pointrange(position = position_dodge( width =
0.75)) +
coord_flip()
```



Estimate for an IQR increase in PFOS when exposure levels to PFOA, PFNA, and PFHxS are fixed at the 25th percentile = 0.16 [-0.08;0.40]

No change in estimate when the other exposures are fixed to the 50th or 75th percentile (no interaction)

BKMR: Pros and cons

Pros

- Capture non-linearity and interaction between exposures
- Can take into account hierarchical structures
  - Reteated exposures
  - Exposure family (grouping individual chemicals by group)

Cons

- Interpretation
- Inference
- Limited number of exposures
- Computationally demanding
- Missing data are not supported
- Convergence

3.

Conclusions

Comparison of statistical approaches

Single-exposure

Multi-exposure

Exposome-wide association study (ExWAS)

To estimate exposure-by-exposure associations

Pros:

- Easy to implement
- High sensitivity
- Complete reporting

Cons:

- High FDR
- Confounding by co-exposure

Variable selection

To identify the set of exposures that predict at best the outcome

Pros:

- Estimates adjusted for co-exposures
- Lower FDR
- Possibility to test interaction

Cons:

- Predictive methods
- Missing data
- Computational resources

Clustering

To identify subjects that share similar exposure pattern

Pros:

- Limit the number of tests
- High dimension and correlation
- Identification of at risk individuals

Cons:

- Interpretation
- Lost of information
- Missing data

Mixture models

To estimate mixture associations and identify interactions between components

Pros:

- Interpretation
- Possibility to test interaction
- Missing data

Cons:

- Limited number of components
- Computational resources

Comparison of methods

Method	Multiple imputed dataset	Non-linearity	Overall effect	Interaction	Computing power	« Force » confounders	Handle categorical exposures	Interpretation
ExWAS	Yes	No	No	Manually	Low	Yes	Yes	Simple
ENET	No	No	No	No	Low	Yes	No	Simple
DSA	Yes	Yes	No	Yes	Medium	Yes	Maybe	Simple
Clustering	No	NA	Yes	Manually	Low	Yes	Yes	Difficult
WQS	Manually	Yes	Yes (one-sided)	No	Low	Yes	No	Simple
Qgcomp	Manually	Yes	Yes	No	Low	Yes	Yes?	Simple
BKMR	Yes (extension)	Yes	Yes	Yes	Large	Yes	No	Complex & mostly graphical

Lessons learned

- Exposome analyses involve **complexity**
- The research question is not always obvious - sometimes multiple
- **Multiple testing**, false positives, negatives and **confounding** are among the main challenges
- Other challenges: missing values, measurement error
- Some methods have been presented to address the most common questions - but this is not an exhaustive list
- We have not covered studies that integrate multiple omics layers, e.g. exposome, transcriptome, metabolome, epigenome,....

Conclusions

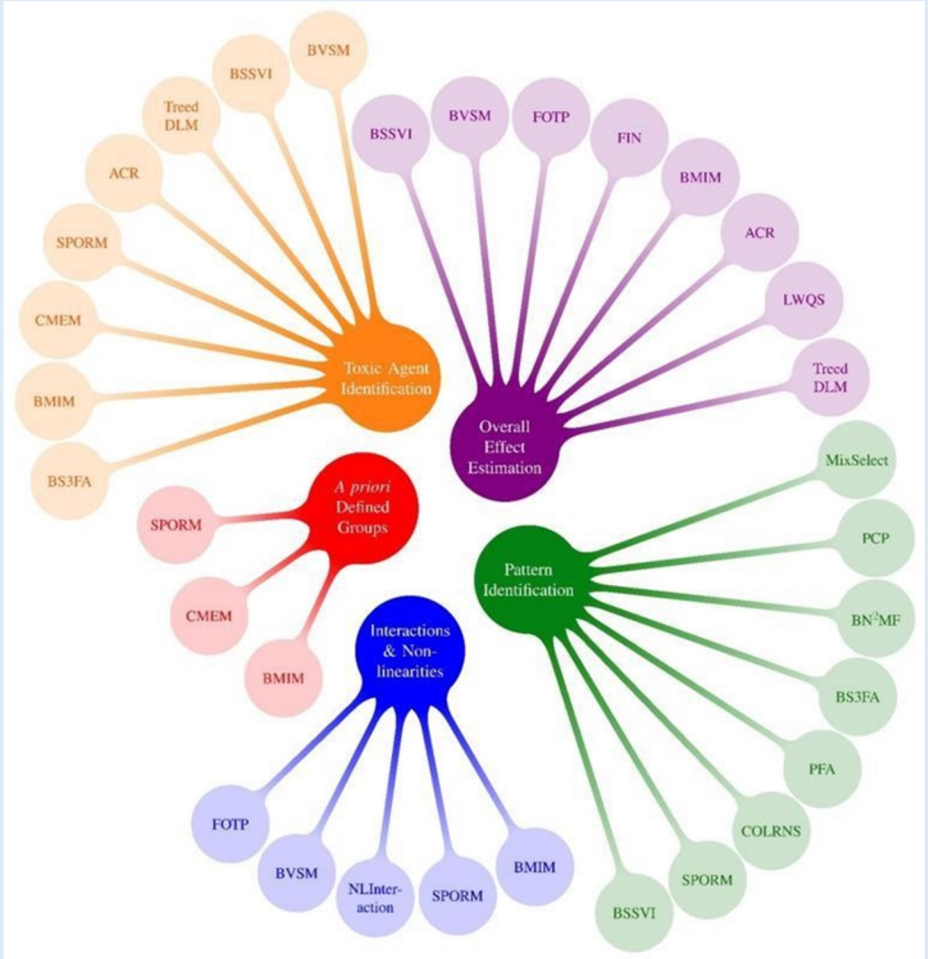
There are no perfect solution

All methods have pros and cons

The choice of method depends on the objective

Several methods may need to be used

The field is in constant evolution



Joubert, Int J Environ Res Public Health 2022

4.

Resources

Exposome data challenge

- Application of several advanced statistical methods to study the exposome

- Summary paper: [Maitre et al. Environ Int 2022]

- Videos of the presentations are available in the ATHLETE project – Exposome Youtube channel:

  https://www.youtube.com/channel/UC0F3hR04UzUeKkcfA yikltA/featured

- Data and analyzes codes are available here:

  https://github.com/isglobal-exposomeHub/ExposomeDataChallenge2021/tree/main/R_Co de_Presentations

References (1/3)

- Agier L, et al. A systematic comparison of linear regression–based statistical methods to assess exposome-health associations. Environmental health perspectives. 2016 Dec;124(12):1848-56.
- Agier L, Slama R, Basagaña X. Relying on repeated biospecimens to reduce the effects of classical-type exposure measurement error in studies linking the exposome to health. Environmental research. 2020 Jul 1;186:109492.
- Barrera-Gómez J, Basagaña X. Models with transformed variables: interpretation and software. Epidemiology 2015 Mar;26(2):e16-7.
- Barrera-Gómez J, et al. A systematic comparison of statistical methods to detect interactions in exposome-health associations. Environmental Health. 2017 Dec;16(1):1-3.
- Brookes ST, et al. Subgroup analyses in randomized trials: risks of subgroup- specific analyses;: power and sample size for the interaction test. Journal of clinical epidemiology. 2004 Mar 1;57(3):229-36.
- Cadiou S, Slama R. Instability of Variable-selection Algorithms Used to Identify True Predictors of an Outcome in Intermediate-dimension Epidemiologic Studies. Epidemiology. 2021 May 1;32(3):402-11.
- Cadiou S et al. Performance of approaches relying on multidimensional intermediary data to decipher causal relationships between the exposome and health: A simulation study under various causal structures. Environment International. 2021 Aug 1;153:106509.
- Chadeau-Hyam M, et al. Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers. Environmental and molecular mutagenesis. 2013 Aug;54(7):542-57.

References (2/3)

- Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015 Aug 14.
- Haug LS et al. In-utero and childhood chemical exposome in six European mother- child cohorts. Environment international. 2018 Dec 1;121:751-63.
- Herzog M.H., Francis G., Clarke A. (2019) The Multiple Testing Problem. In: Understanding Statistics and Experimental Design. Learning Materials in Biosciences. Springer, Cham. https://doi.org/10.1007/978-3-030-03499-3_5
- Lazarevic N, Knibbs LD, Sly PD, Barnett AG. Performance of variable and function selection methods for estimating the nonlinear health effects of correlated chemical mixtures: A simulation study. Statistics in Medicine. 2020 Nov 30;39(27):3947-67.
- MacLehose RF, Dunson DB, Herring AH, Hoppin JA. Bayesian methods for highly correlated exposure data. Epidemiology. 2007 Mar 1:199-207.
- McGee G, Wilson A, Webster TF, Coull BA. Bayesian Multiple Index Models for Environmental Mixtures. https://arxiv.org/abs/2101.05352
- Papathomas M, et al. Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers. Environmental health perspectives. 2011 Jan;119(1):84-91.
- Patel CJ, Manrai AK. Development of exposome correlation globes to map out environment-wide associations. InPacific Symposium on Biocomputing Co-Chairs 2014 (pp. 231-242).

References (3/3)

- Robinson O, et al. The pregnancy exposome: multiple environmental exposures in the INMA-Sabadell birth cohort. Environmental science & technology. 2015 Sep 1;49(17):10632-41.
- Robinson O, et al. The urban exposome during pregnancy and its socioeconomic determinants. Environmental health perspectives. 2018 Jul 17;126(7):077005.
- Santos S, Maitre L, Warembourg C, Agier L, Richiardi L, Basagaña X, Vrijheid M. Applying the exposome concept in birth cohort research: a review of statistical approaches. European Journal of Epidemiology. 2020;35(3):193-204.
- Stafoggia M, Breitner S, Hampel R, Basagaña X. Statistical approaches to address multi-pollutant mixtures and multiple exposures: the state of the science. Current environmental health reports. 2017 Dec;4(4):481-90.
- Tamayo-Uria I, et al.. The early-life exposome: description and patterns in six European countries. Environment international. 2019 Feb 1;123:189-200.
- Vineis P, van Veldhoven K, Chadeau-Hyam M, Athersuch TJ. Advancing the application of omics-based biomarkers in environmental epidemiology. Environmental and molecular mutagenesis. 2013 Aug;54(7):461-7.
- Warembourg C, et al. Statistical Approaches to Study Exposome-Health Associations in the Context of Repeated Exposure Data: A Simulation Study. Environ Sci Technol. 2023 Oct 31;57(43):16232-16243.
- Warembourg C, et al. Urban environment during early-life and blood pressure in young children. Environment International. 2021 Jan 1;146:106174.
- Witte JS, et al. Multilevel modeling in epidemiology with GLIMMIX. Epidemiology. 2000 Nov 1;11(6):684-8.

Software

- rexposome R packages implements many of these analyses.
- TO tutorial es available online:

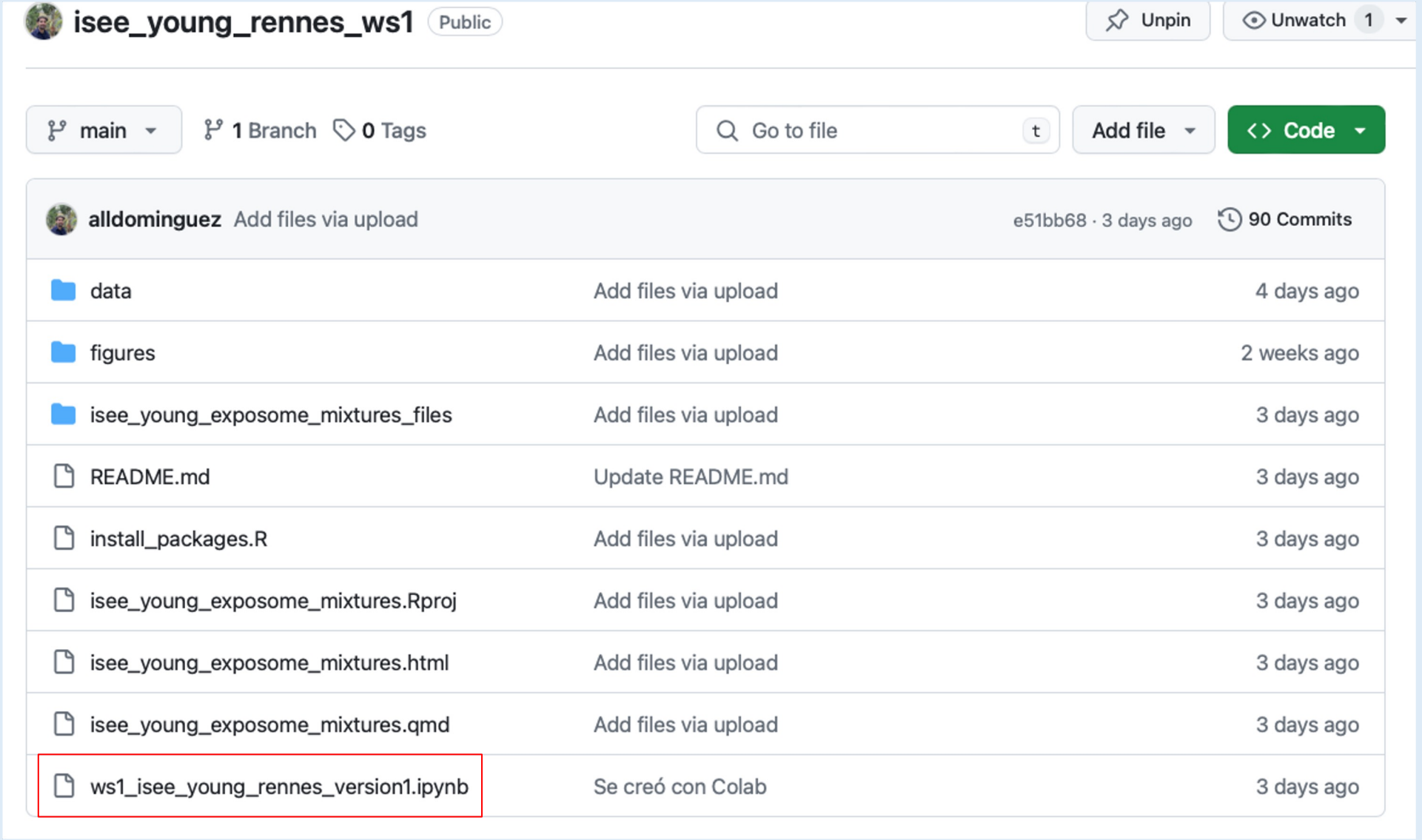
https://rpubs.com/jrgonzalezISGlobal/rexposome

Thank you

Hands-on session

https://github.com/alldominguez/isee_young_rennes_ws1

Google collab - Option 1



Option 1 - Google collab

isee_young_rennes_ws1 / ws1_isee_young_rennes_version1.ipynb

alldominguez Se creó con Colab — 9e27cea · 3 days ago — History

Preview | Code | Blame | 3735 lines (3735 loc) · 738 KB | Code 55% faster with GitHub Copilot | Raw

Open in Colab

Workshop 1: Statistical methods for studying mixtures and the exposome

The study of mixtures and the exposome in the context of environmental epidemiological research is rapidly growing. Investigating mixtures and the exposome allows researchers to assess the independent and combined effects of various exposures, as well as their potential synergistic or antagonistic effects, on health outcomes. However, the complexity of exploring these questions requires the use of specific statistical models to account for aspects that single-exposure models cannot typically handle (e.g. multicollinearity).

This workshop therefore aims at summarizing and presenting the main models used for studying mixtures and the exposome, and discussing the pros and cons of each method in relation to a specific study objectives.

Introduction to the NoteBook 📚

Option 1 - Google collab

ws1_isee_young_rennes_version1.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

+ Código + Texto Copiar en Drive

Workshop 1: Statistical methods for studying mixtures and the exposome

The study of mixtures and the exposome in the context of environmental epidemiological research is rapidly growing. Investigating mixtures and the exposome allows researchers to assess the independent and combined effects of various exposures, as well as their potential synergistic or antagonistic effects, on health outcomes. However, the complexity of exploring these questions requires the use of specific statistical models to account for aspects that single-exposure models cannot typically handle (e.g. multicollinearity).

This workshop therefore aims at summarizing and presenting the main models used for studying mixtures and the exposome, and discussing the pros and cons of each method in relation to a specific study objectives.

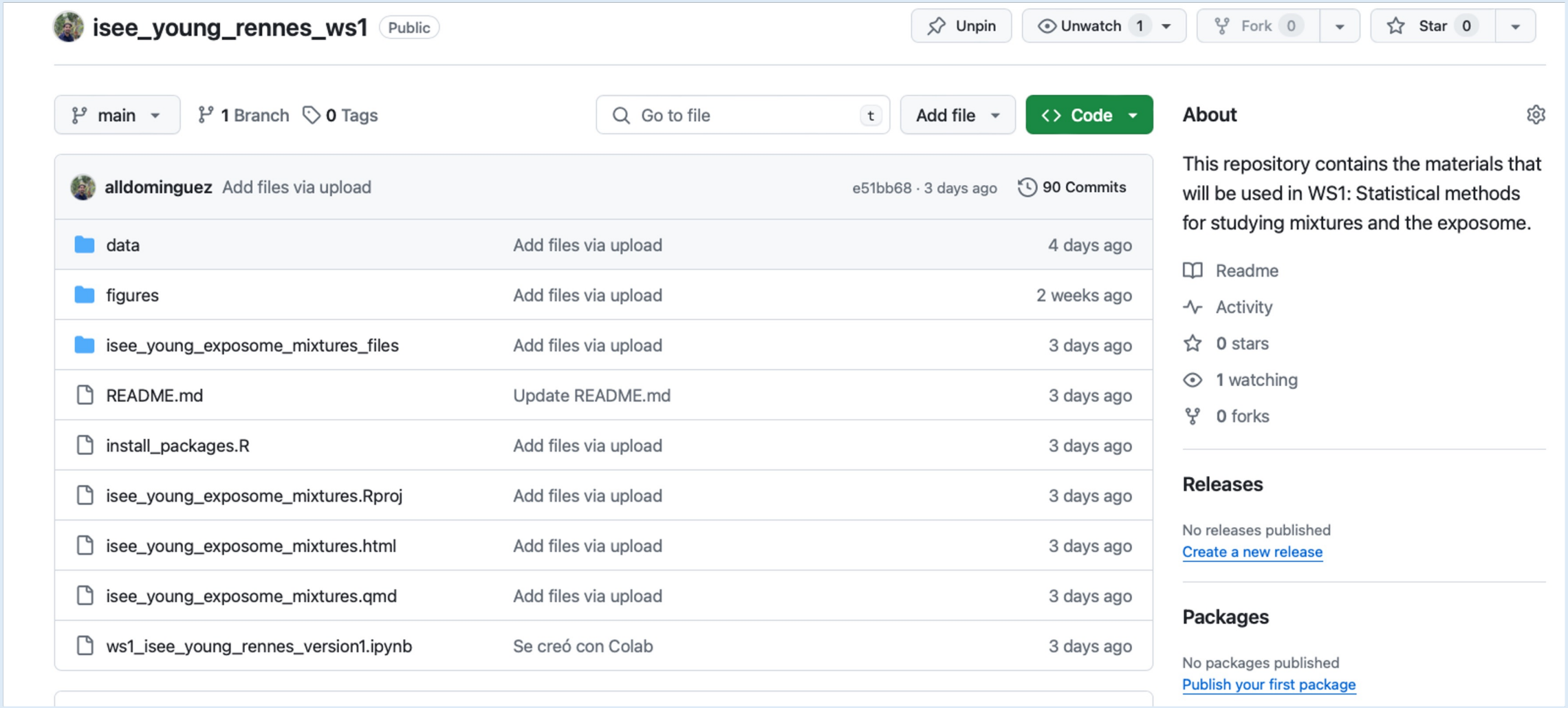
Introduction to the NoteBook 📚

Within this **NoteBook**, you will be guided step by step from loading a dataset to running some mixture and exposome analysis.

The Jupyter notebook is an interactive computing environment that allows users to author notebook documents. Notebooks consist of **linear sequence of cells** that combines **code cells (input and output of live code that is run)**, and **markdown cells (narative text)**.

The components of the notebook are:

Option 2 - Rstudio (Quarto)



Option 2 - Rstudio (Quarto)