HBM4EU project

Statistical analyses of HBM data in HBM4EU

Eva Govarts - VITO

2nd HBM4EU Training School 2018

HBM4EU data analysis

Aim: data analysis at EU level

Needs:

- Harmonized variables
- Central data analysis

→ harmonized codebook developed by WP10

→ research protocols developed by partners

HBM4EU data analysis

Data: biomarker (HBM) and questionnaire data

cadmium

bisphenols

phthalate metabolites

mercury

HBM4EU statistical analysis plan (task 10.3)

- Statistical issues common to all HBM analysis
- Descriptive statistics
- Some typical statistical analyses
  - Time trends
  - Geographical comparisons
  - Exposure determinants
- Substance specific statistical analysis
- Calculation of European reference values

Statistical common issues

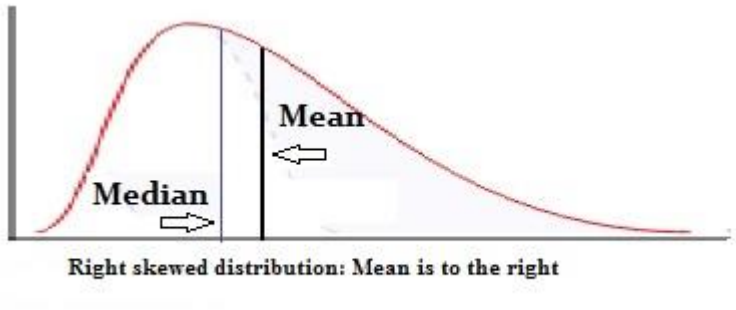
HBM variables = biomarker data

- Transformation
- Values below LOD or LOQ
- Missing data
- Adjustment
  - Urinary HBM analysis
  - Blood HBM analysis

Transformation

Biomarker values

- Typically right-skewed
  - Solution: ln-transformation



Right skewed distribution: Mean is to the right

- Other distributions possible
  - Solution: categorization

In theory: normality not a requirement in regression models!!

→ however, normality of residuals is, and transformation helps

Values below LOD or LOQ

Two laboratory limits

- Limit of detection = LOD

  → lowest analyte concentration at which detection is feasible

- Limit of quantification = LOQ

  → lowest analyte concentration at which detection is reliable (predefined goals for bias and imprecision)

Definition and determination depends on the lab.

⟹ should be compared for existing data collections

- ✓ definition/determination
- ✓ value of LOD and LOQ
- ✓ % values < LOD or LOQ

Values below LOD or LOQ

Handling values below LOD or LOQ

- Complete case analysis (CCA)

  → elimination of values → bias

- Replacement by fixed value

  e.g. LOD/2 or LOQ/2

- Single Imputation

  random selection of possible values (between 0 and LOD) from a log-normal probability distribution, dependent on the study population and observed values for other contaminants, and allowing residual variances to vary by population

- Multiple imputation

  similar to single imputation, but imputes more than one value, resulting into multiple datasets

Values below LOD or LOQ

Recommendation for HBM4EU

Depending on %<LOD or LOQ

- >80% → dichotomize variable
- Otherwise: single imputation

Missing data

Both for HBM variables as questionnaire variables missing values typically occur

- HBM variables: complete case analysis advised (if missing at random!)
- Questionnaire variables: advised there to use (multiple) imputation

Adjustment

Biomarkers measured in

- Urine
  - Adjustment for dilution level exposure biomarkers
- Blood
  - Adjustment for blood lipids for fat-soluble exposure biomarkers

Urinary dilution

Urinary dilution measures

- Creatinine
  → **creatinine excretion (g/L)**
- Osmolality
  → **concentration of a solution expressed as Osm/kg**
- Specific gravity
  → **ratio of the density of urine to the density of distilled water**

In existing data collections: limited to available measures!!

Methods to adjust or normalize for urinary dilution

- The ratio model
- Normalization to a standard urinary concentration
- The covariate model

Adjustment for blood lipids for fat soluble exposure biomarkers

- Adjustment by blood lipid level

  ng substance / ng lipids

- Add lipid to model

Questionnaire variables

- Missing values: imputation

- Calculating new variables
  - Recoding categories
  - Combining variables
  - Calculations
  - Categorising continuous variables

Descriptive statistics

Aim: picture of study population

- Per data collection
- Pooled database

Aim: analysis of biomarkers

- Descriptive statistics
- Stratification by possible determinants of exposure
  - ANOVA
  - Multiple linear regression model (see later)

Study population

Descriptive statistics

Variables: all covariates used in further analyses + general parameters

- Participant characteristics: age, sex, smoking, alcohol, BMI, etc.
- General characteristics: housing condition, neighboring industry, occupation, etc.
- Sample conditions: urinary volume, hair structure, etc.
- Toxicological parameters: creatinine, SG, lipids, etc.

Calculations

- Continuous variables
  - N
  - Median (P25-P75)
  - Min-Max

- Dichotomous/categorical variables
  - N (%)

Example of table format

Mothers (N = xxx)			
Parameter	Statistics	Values	
Age, years	Total N	xxx	
	Median	xx	
	P25 – P75	xx	xx
	Min.-max.	xx	xx
Age distribution:	Total N	xxx	
<30 years	N, %	xx	xx.x%
30-40 years	N, %	xx	xx.x%
≥40 years	N, %	xx	xx.x%
Urinary creatinine, mg/L	Total N	xxx	
	Median	xx	
	P25 – P75	xx	xx
	Min.-max.	xx	xx
Urinary creatinine in classes	Total N	xxx	
< 300 mg/L	N, %	xx	xx.x%
300 – 1000 mg/L	N, %	xx	xx.x%
1000 – 1500 mg/L	N, %	xx	xx.x%
1500 – 3000 mg/L	N, %	xx	xx.x%
> 3000 mg/L	N, %	xx	xx.x%

Biomarker results

- Results for total group
- Stratified results by relevant subgroups
  - Relevant subgroups, e.g. age class, sex, ISCED, etc.
  - Selection of covariates

start from toxicokinetics of biomarker

select relevant covariates

- Comparison between data collections / countries

Prerequisite for valid analysis → each cell should contain ≥ 5 observations

THEREFORE: some variables may need recoding:

- Combining categories (in case of more than two subgroups):

  e.g. smoking status: less than 5 occasional smokers → combine occasional & daily smokers

- Redefining categories (in case of two subgroups): change definition

  e.g. food consumption: e.g. less than 5 participants eat hazelnut spread several times per week → redefine subgroups to at least once per month

- Skipping variables (in case of yes/no variable): too few observations to allow analysis

  e.g. industry neighborhood of residence: if less than 5 participants live close to industry, this variable is not analysed further

Calculations

- N
- N > LOD or LOQ
- Percentiles: P5, P10, P25, P50, P75, P90, P95
- Arithmetic mean + standard deviation (SD)
- Geometric mean + 95% confidence interval

Example of table format

Age group: mothers, 18-45 years
Biomarker: mercury in hair
Unit: µg/mg
LOQ: xx.xxx µg/mg

Strata	N	% >LOQ	GM	low CI	up CI	AM	SD	min.	P10	P25	P50	P75	P90	P95	max
All	xxx	xx.x%	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xx
Age (p=0.xx)															
<30 years	xxx	xx.x%	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xx
30-40 years	xxx	xx.x%	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xx
≥ 40 years	xxx	xx.x%	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xxx	xx.xx

Typical statistical analyses

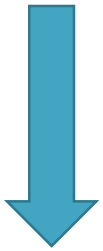
Exposure related policy questions related to

- Time trends
  - Are there different time trends for unregulated and regulated phthalates?
  - Is there a sufficient decrease of BPA exposure after the European regulation?
  - …
- Geographical comparison
  - Can countries be grouped according to similarity in concentration levels?
  - What is the impact of EC restriction of manufacturing, marketing and use of PFOA under REACH?
  - …
- Exposure determinants
  - What are determinants of concentration levels?
  - Is exposure driven by diet, consumer exposure, occupation or environmental contamination?
  - …

Exposure determinants — *Typical statistical analyses*

Univariate analysis → considers only 1 factor at the time

→ possible errors in interpretation due to co-linearity



Multiple regression models

→ include in model: pre-defined confounders + covariates

→ final model gives relative contribution of different determinants at the same time

→ partial R^2: % of variability explained

→ beta coefficients: strength of association between biomarker and factor in a model, after adjustment for all other factors in the model

Time trends

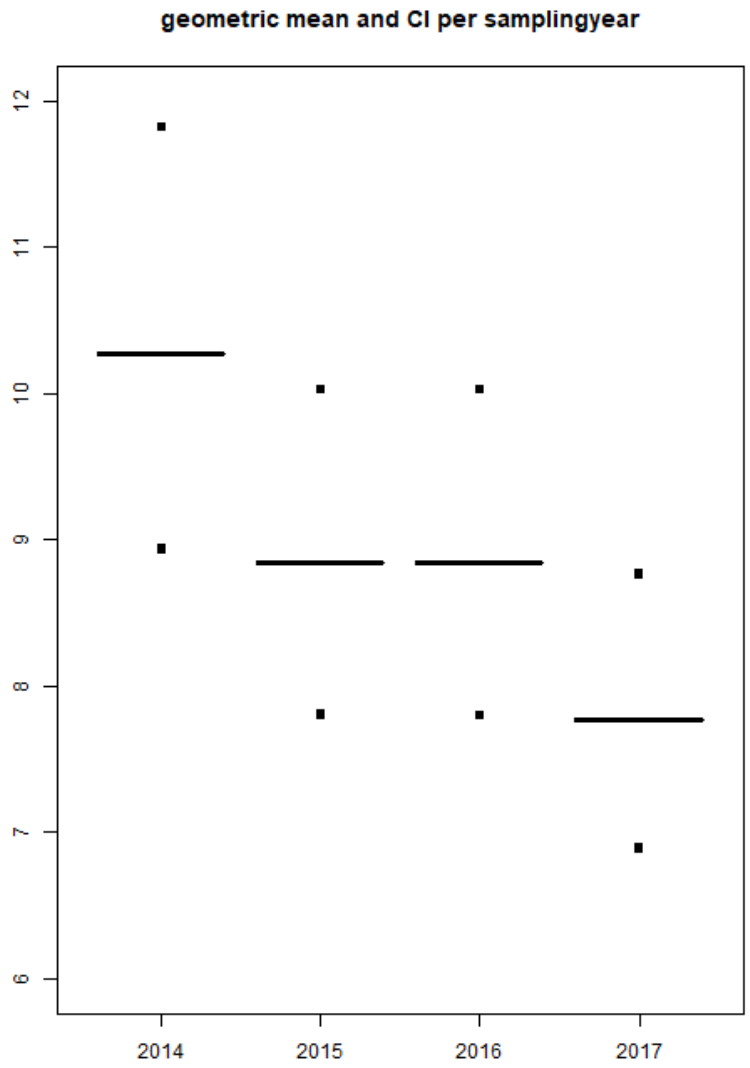
Typical statistical analyses

Descriptive analysis

- Descriptive statistics
- Descriptive plots

Statistical analysis

- T-test/ANOVA
- Regression model
- MANOVA
- Mixed models
- Generalized estimating equations (GEE)



Geographical comparison *Typical statistical analyses*

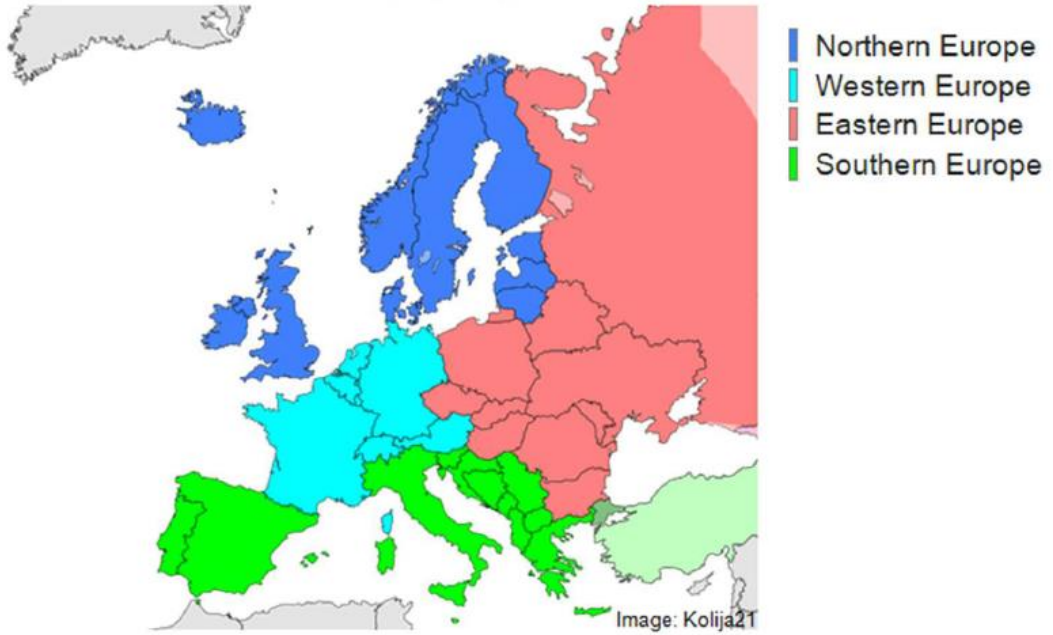
Descriptive analysis

- Descriptive statistics
- Descriptive plots

Statistical analysis

- T-test / ANOVA
- Regression model

Graphic: Subregions of Europe (UN geoscheme)



Substance specific

1st statistical analysis plan

- General part
- Substance specific part on 1st set of prioritized substances

  = Deliverable D10.2 (December 2017)

- Substance specific part on 2nd set of prioritized substances

  = Deliverable D10.5 (February 2019)

Describes

- Exposure related research questions as formulated in scoping documents

- List of variables needed to solve these research questions

Research protocols

For each specific substance

- Formulate research question of interest
- Determine variables needed to explore research question
  - Harmonized variables of codebook
  - Additional variables: proposal for harmonization (input to WP10)
- Develop research protocol
- Identify suited data collections
  - Specify if individual or aggregated data are needed
- Send them research protocol
  - If already contacted by WP10, additional motivation to speed up process of data sharing
  - If not and interested, explanation steps of data sharing by WP10
- Access to the data via HBM4EU repository (after signing agreements)
- Statistical analysis in preferred software (R, SAS, SPSS, …)
- Assistance by WP10 statistical working group

→ PUBLICATION ☺

Research protocols

Exposure related research protocols under development (task 10.4)

- Time trends in Cat A & B **phthalates** in Danish and German young adults between 2000 and 2017
- Urinary levels of **BPA** among European women and major determinants
- Geographic variations in Cat A **Flame retardants** in the European population
- **Cadmium** exposure among European residents and its geographical variability
- **PFASs** levels and determinants of exposure in vulnerable population groups
- Exposure determinants of Cat A **phthalates**

- Exposure distributions – European Reference Values (ERVs): phthalates/DINCH, BPA, PFASs

Thank you for your attention

If you have any questions about data analysis

Don’t hesitate to contact the HBM4EU WP10 statistical working group!

VITO: eva.govarts@vito.be

ISGlobal: martine.vrijheid@isglobal.org

UBA: nina.vogel@uba.de

ANSP: loic.rambaud@santepubliquefrance.fr

Contacts

Eva Govarts: eva.govarts@vito.be
Sylvie Remy: sylvie.remy@vito.be
Liese Gilles: liese.gilles@vito.be
Greet Schoeters: greet.schoeters@vito.be

Speaker's information

Eva Govarts works as researcher – biostatistician at the Flemish Institute for Technological Research (VITO), Mol, Belgium. She received training in biomedical sciences, applied and biostatistics. In HBM4EU she is task leader of task 10.4 on the data analysis and generation of European reference values (RVs) and together with Greet Schoeters she is co-leading WP10.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733032.