UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Derivation of European exposure values of internal exposure to environmental pollutants using human biomonitoring data

Paraskevi Filippousi

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :
dr. Liesbeth BRUCKERS

SUPERVISOR :
Mevrouw Eva GOVARTS

2024
2025

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Derivation of European exposure values of internal exposure to environmental pollutants using human biomonitoring data

Paraskevi Filippousi

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :

dr. Liesbeth BRUCKERS

SUPERVISOR :

Mevrouw Eva GOVARTS

Abstract

Human biomonitoring, measuring chemicals or their metabolites directly in tissues and fluids, can, in principle, reveal EU-wide exposure patterns. The pooled HBM4EU data (2014–2021) were assembled from national and regional cohorts that each followed different, often not clearly documented, sampling schemes. The resulting dataset lacks a unified probabilistic design, and any uneven coverage against geographic and socio-demographic aspects, as well as the absence of sampling weights make the derivation of "European" reference exposure values a challenge. This thesis focuses on the HBM4EU children's age-group (6–12 yrs) and phthalates/mono-benzyl phthalate (MBzP) as a test-case. Exploratory analysis of the children dataset confirmed a North–East bias, an excess of high-education households and urban–rural mismatches; sampling year shadows cohort, magnifying site heterogeneity. Pronounced MBzP gradients by region, DEGURBA, sampling season and education affirmed the need for weights and cluster-robust inference, potentially providing a transferable template for other future initiatives.

A population–standardisation grid for EU-27 children was built crossing one-way Eurostat margins for *region* (North, South, West, East), *sex* (male, female), *season* (each pre-weighted at 0.25), *DEGURBA* (urban, towns/suburbs, rural) and *household-education* (ISCED 0–2, 3–4, ≥5). Age was fixed at 9 years, considering also the regional Eurostat data showing that uniform single-year counts across the 6–12-yr span. The Cartesian product yields 288 cells; each cell weight equals the product of its five marginal proportions and the set is normalised to 1. This construction assumes the five dimensions are mutually independent; in the absence of joint tabulations. These grid weights served a dual role: for **model–based routes:** each regression was fitted to the HBM4EU children data, after which its fitted values were projected onto the 288 cell profiles and post-stratified with the grid weights to represent an average EU child. Two specifications were considered: (i) an ordinary-least-squares model, with and without region–specific interaction blocks, evaluated with delta-method SEs; and (ii) a random-intercept mixed model, with and without interactions, propagating uncertainty via the analytic Delta-method, as well a "MC-fixed" Monte-Carlo SE (resampling) only the fixed-effect coefficients but also a "MC-full" Monte-Carlo SE (resampling both the fixed effects and a new cohort-level intercept on every replicate). For **design–based routes:** the same weights were merged back to the HBM4EU data; dividing each cell weight by the number of sampled children in that cell, yielding observation–level probabilities that drove direct post-stratification, survey-design analysis and marginal raking, with weight trimming explored as sensitivity checks. The EU-27 standardised geometric mean estimates yielded narrower ranges, both with the model-based and design-based approaches, once cohort clustering was not considered. Declaring cohorts as PSUs inflated the confidence bands. Weight-trimming and marginal raking reduce design effects and sharpen intervals with negligible impact on the central estimate, whereas extending the calibration to a region×age margin lowered the geometric mean notably while substantially cutting the effective sample size. Next steps could focus on: (a) future initiatives collecting study-specific design weights *before* pooling to keep all downstream estimates design-consistent; (b) reporting both an efficient mixed-effects projection (MC-fixed SE) and a raked, cluster-robust survey estimate to bracket uncertainty; (c) test on the effect of replacing regional margins with finer, e.g. country-level or joint margins (if available); (d) include single-year age calibration and extend the framework to adolescents, adults and further biomarkers; and (e) explore weight- and cluster-aware design-based regression (svyglm) or GEE/GEE2, providing population-average estimates with sandwich-robust SEs.

Keywords: *HBM4EU; biomonitoring; biomarkers; chemicals exposure; phthalates; weighting; direct standardisation; post-stratification.*

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Prof. Liesbeth Bruckers for her support, constant encouragement, and scientific guidance throughout this challenging and complex project. My thanks also go to my VITO advisors, Dr. Eva Govarts and Dr. Hamid Hassen, for their expertise/guidance around the HBM4EU data set and related prior research, as well as all the discussions on results and next steps. Thanks also to Dr. Bianca Cox for her review and constructive feedback on part of the thesis.

I am profoundly grateful to my family and friends for their patience and unwavering encouragement while I tried to balance full-time work with this master's programme. Our time together and simply them being there—even on the difficult days—brought me within sight of the finish line.

Finally, I wish to thank Dr. Vercruysse and Ms. Thijs for their reliable administrative support, and Mr. Vandebempt for his ever-positive willingness to help with all the practical challenges faced by a working student.

1 Introduction

1.1 Human biomonitoring

Human biomonitoring (HBM) measures the concentration of chemicals or their metabolites in human fluids and tissues. Therefore, it allows the assessment of "human exposure to chemicals from different sources, by different pathways, and during certain periods of life".[1] It thus provides a comprehensive assessment of overall exposure by capturing chemicals intake from a variety of sources, including environmental, occupational, dietary, and consumer products. The European Commission's Chemicals Strategy for Sustainability (CSS), published in 2020, explicitly recognizes human biomonitoring (HBM) as a vital tool for assessing chemical exposure and informing policy decisions.[2] The CSS highlights the importance of HBM in understanding the internal concentrations of chemicals, thus supporting the EU's zero-pollution ambition. Furthermore, CSS outlines the development of a framework of indicators to monitor the drivers and impacts of chemical pollution, including the use of HBM data to measure the effectiveness of chemicals legislation.[3]

1.2 The European Human Biomonitoring Initiative (HBM4EU)

HBM4EU, launched in 2017, aspired to advance and harmonise human-biomonitoring efforts across Europe.[5] It encompassed national and regional studies, building on existing capacity of countries with recurring HBM programs, such as Germany, Belgium (Flanders), France, Sweden, and Slovenia, while 32% of the studies were initiated specifically under the guideline protocols of the HBM4EU project (e.g. for Greece, Portugal, Croatia, Switzerland, and others).[6] [4] Consequently, this suggests that each participating study retained its own sampling frame and recruitment strategy and no single EU-level probabilistic design underpins the pooled dataset and statistical representativeness is guaranteed only within the boundaries of each contributing survey. Participants were recruited between 2014 and 2021, from approximately 11–12 countries per age group, ensuring a broad representation across Europe. The studies included 10,795 participants in three age groups: 3,576 children (6–12 years), 3,117 teenagers (12-18 years) and 4,102 young adults (20-39 years). Each participating study (primary sampling unit, PSU) followed a common HBM4EU-aligned protocol: for instance, within the 6–12-year stratum ca. 300 children were retained per PSU, with sex quotas of $\approx 50\,\%$ boys and $\approx 50\,\%$ girls and with $\geq 10\,\%$ representation in every DEGURBA class (urban, towns/suburbs, rural) and in each household-education level (ISCED 0–2, 3–4, ≥ 5). The twelve PSUs under the children's age-group were allocated across the four UN geoscheme regions (North, South, West, East) and sampling spanned all four seasons.

HBM4EU samples were analyzed for specific biomarkers, indicative of exposure to various chemical substances. Those included a range of chemicals, including emerging contaminants and legacy pollutants, to better understand population exposure levels and trends. A prioritisation strategy led to the identification of a list of priority substances for HBM in Europe.[7] The first list of high-priority substances for action in HBM4EU included 9 substance groups: 1) phthalates and the phthalate alternative: DINCH, 2) bisphenols, 3) per- and polyfluoroalkyl substances (PFAs), 4) (organophosphorus and halogenated) flame retardants, 5) cadmium and chromium VI, 6) polycyclic aromatic hydrocarbons (PAHs), 7) aromatic amines, 8) chemical mixtures and 9) emerging substances. The second list expanded into: acrylamide, aprotic solvents, arsenic, diisocyanates,

lead, mercury, mycotoxins, pesticides, and benzophenones. An HBM4EU expert group selected the most appropriate biomarkers for each of the priority substances.[8] For the HBM4EU aligned studies, a relatively large sample size was attained for the different substance groups being measured. For example, with regard to children: pesticides (6 countries, N=863), phthalates/DINCH (12 countries, N=2877), organophosphorus flame retardants (7 countries, N=1,768), halogenated flame retardants (4 countries, N=710) and acrylamide (5 countries, N=1,198).[6]

1.3 Context and thesis research question

Table 1 summarises information identified under the "Information Platform for Chemical Monitoring" (IPCHEM), but also under published supporting material, related to the characteristics of the HBM4EU-related cohorts, encompassing the children (6-12 years) age-group.[9],[4] One can conclude that cohorts were recruited with different, and in most cases, not clearly-documented sampling strategies which ranged from national, stratified surveys to regional convenience or hospital-based birth cohorts. Unequal coverage of countries, regions, age-groups, socio-economic strata and exposure settings means that some demographic profiles may be over- or under-represented. One solution for "European"-level representative estimates could be based on a hybrid scheme: explicit design weights where the design is known, and externally-derived calibration weights where it is not. Since study sampling designs are not clearly documented or heterogeneous, post-stratification weights that rely on known selection probabilities may become unreliable, suggesting the need to look for external standards (e.g. Eurostat) for applying weights related to socio-demographic and geographic variables.

Table 1: HBM4EU: summary of children-related cohort characteristics

HBM4EU Children cohorts (Sampling year)	**Country**	**N**	**Sampling strategy**	**Age**	**Study design**
InAirQ (2017-2018)	Hungary	262	IPCHEM: Probabilistic Supporting Info rather indicates the possibility of convenience sampling	8-11	Cross-sectional
NACII (2014-2016)	Italy	300	IPCHEM: Probabilistic Supporting Info rather indicates purposive sampling and convenience sampling as a secondary aspect	7 only	Cross-sectional
GerESV (children subset) (2015-2017)	Germany	300	Stratified random	3-17	Cross-sectional
NEBII (children subset) (2016-2017)	Norway	300	IPCHEM: undefined Supporting Info indicates non-random, purposive (or targeted) sampling.	7-11	Longitudinal

ESTEBAN (children subset) (2014-2016)	France	543	IPCHEM: undefined Supporting Info indicates two-stage sampling: random household selection with individual-level exclusions	6-17	Cross-sectional
POLAES (children subset) (2017)	Poland	300	IPCHEM: undefined Supporting Info indicates mixed convenience and purposive sampling	7-10	Case-control
PCB (children subset) (2014-2017)	Slovakia	300	IPCHEM: maternal approval at delivery Supporting Info indicates clinically-based convenience sampling	10-12	Longitudinal
SLOCRP (children subset) (2018)	Slovenia	149	IPCHEM: Undefined Supporting Info indicates convenience sampling with purposive exclusions	7-10	Cross-sectional
CROME (children subset) (2020-2021)	Greece	161	IPCHEM: Simple random Supporting Info indicates convenience and snowball sampling	6-11	Cross-sectional
OCC (children subset) (2018-2019)	Denmark	300	IPCHEM: Random selection - stratified Supporting Info rather indicates longitudinal birth cohort sampling	5-7	Longitudinal
3xG (children subset) (2019-2020)	Belgium	133	Hospital-based prospective birth cohort	6-8	Longitudinal

The overarching aim of this thesis has been to evaluate and compare statistical approaches for deriving EU-level, reference values of chemical exposure, from biomarker data generated by the studies included under HBM4EU. In HBM4EU, weights have not been used for the calculation of European exposure values (or other analyses); thereby the use of weights may result in more EU-level representative values. Assessing different strategies/methodologies could be useful for (future) calculations of EU-exposure values, e.g. based on the PARC aligned studies.

Since simultaneously addressing all age strata and the full set of biomarkers would be very complex at once, and considering that the purpose has been to obtain methodological insights, the present work focuses on (i) the **children** age-group (6–12 years) and (ii) **phthalates**, with the example the phthalate metabolite **mono-benzyl phthalate (mbzp)**, a primary urinary biomarker of the plasticiser butyl benzyl phthalate (BBzP). BBzP has historically been added to flexible PVC flooring, sealants, coated fabrics and other building or consumer materials; legacy uses make it a continu-

ing indoor source despite recent regulatory restrictions. Methodologically, the study couples the HBM4EU dataset with externally-sourced population margins, obtained from Eurostat, in order to build a weighting scheme representing the EU-27 demographic mix of 6–12-year-olds and deliver population-standardised means and standard errors via selected statistical methodologies.

1.4 Societal relevance, Stakeholders and Ethics

Human Biomonitoring for Europe (HBM4EU) plays a key role in assessing the efficacy of chemicals management policies and the monitoring of emerging pollutants. By systematically measuring internal exposure to hazardous substances, regulatory frameworks such as REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) could be informed and the EU could make bigger steps towards its Zero-pollution ambition. HBM4EU has pooled data from a network of coordinated national studies. Establishing EU-level reference values can be crucial as they can provide a common baseline against which Member States can evaluate national results, highlighting regions where exposures are atypically high or low. Moreover, it could enable proportionate, evidence-based policy targets that are coherent across borders but also enable the timely detection of emerging chemical hazards across Europe. Finally, EU-reference values could provide a robust foundation for establishing benchmarks for comparison with chemicals exposure data worldwide.

HBM4EU outcomes are of relevance to a broad spectrum of stakeholders. Policy makers and regulatory bodies at EU, national, and regional levels could use relevant findings to calibrate exposure limits and prioritise risk-mitigation strategies. Furthermore, public health and environmental agencies may use the data to target interventions in high-risk areas. Civil society, encompassing citizens, consumer associations, and non-governmental organisations (NGOs), can benefit from transparent reporting of chemical exposures and flagging potential health implications across the EU.

Fully anonymised datasets were extracted from the PEH Data Platform (Personal Exposure and Health Data Platform) and provided by VITO. The individual-level data were fully anonymised to protect participant confidentiality. Moreover, every aligned study has obtained approval from its relevant local ethical committee, and explicit consent procedures have been implemented for the sharing of personal data at the European level. Further information with regards to HBM4EU cohorts meeting ethics requirements and details on local ethical committees has also been previously summarised.[6] These measures protect participants' rights and ensure research integrity.

2 Data

2.1 Data collection

A subject-coded, fully anonymised dataset was applied hereby, covering the age-group of children (6-12 yrs). The HBM4EU dataset for children contained in total 2823 entries. It included biomarkers related to two pollutant categories: i) Phthalates (concentrations for 14 biomarkers, as well as 10 sum parameters) and ii) Flame retardants (concentrations for 2 biomarkers). Creatinine (crt) was used as an estimator for urinary density and a parameter to standardise the biomarker concentrations. More specifically, the study hereby used biomarker concentrations expressed as biomarker_impcrtlog, namely the natural–logarithm of the imputed biomarker concentration, standardised to creatinine (expressed as μg g^{-1} crt). This already covered imputation for values below the Limit of Detection (LOD) and Limit of Quantification (LOQ). An additional overview of the full set of specific biomarkers within the children dataset is summarised, only for information, under the Appendix (section B: additional tables).

2.2 Dataset dictionary

The list of all variables included in the provided HBM4EU dataset (children 6-12 years age-group), and their key characteristics, is summarized, for reference, in Table 2.
Variables retained for subsequent statistical analyses are shown, within the table, marked by double *.

Table 2: Summary of variables included in the extracted HBM4EU children's dataset

Variable	Description	Type	Values / codes
`cohort_name`	11 HBM cohorts	string	1=C_NPHI_InAirQ 2=C_EPIUD_NAC II 3=C_UBA_GerES V 4=C_NIPH_NEB II 5=C_ANSP_ESTEBAN 6=C_NIOM_POLAES 7=C_SZU_PCB cohort 8=C_JSI_SLO CRP 9=C_AUTH_CROME 10=C_SDU_OCC 40=C_VITO_3xG
`*cohort*`	cohort code	integer	1–10, 40 (see above)
`country`	country related to the cohort	string	HU, IT, DE, NO, FR, PL, SK, SL, EL, DK, BE

`nuts1`
`nuts2`
`nuts3`
	NUTS level of participant's residence:
	NUTS 1: major socio-economic regions
	NUTS 2: basic regions
	(for regional policies)
	NUTS 3: small regions
	(for specific diagnoses)
		alphanumeric string	official Eurostat codes (e.g. BE2, BE24, BE241)

region	geographical region (UN geoscheme)	integer	1=North (DK, NO) 2=South (SI, GR, IT, CY) 3=West (FR, NL, DE, BE) 4=East (PL, HU, SK)
id_hbm4eu_subject	unique participant ID	alphanumeric string	agegroup_institution_study_ID, e.g. T_VITO_FLEHSIV_1
matrix	biological matrix	string	US=urine-spot, UM=urine first-morning
crt	concentration of creatinine in urine samples	numeric	µg/L
samplingyear	year of sample collection	integer	2014–2021
samplingmonth	month of sample collection	integer	1–12
samplingday	day of sample collection	integer	1–31
samplingtime	time of day of sampling	integer	1= morning, 2= afternoon, 3= evening, 4= night
samplingseason	season of sampling	integer	1= spring, 2= summer, 3= autumn, 4= winter
sex	sex of the participant	character	M= male, F= female
height	height at sampling	numeric	cm
weight	weight at sampling	numeric	kg
bmi	body-mass index	numeric	kg/m^2
ageyears	age in years at sampling	integer	6–12
agemonths	age in months at sampling	integer	72–156
smoking_passive	passive smoking exposure at home	integer	0 = no, 1 = yes
degurba	degree of urbanisation	integer	1 = cities; 2 = towns/suburbs; 3 = rural
isced_hh	highest education level of the household of the subject at sampling (ISCED scale)	integer	1=Low (ISCED 0–2), 2=Medium (ISCED 3–4), 3=High (ISCED ≥ 5)

biomarker	biomarker (e.g.mbzp) concentration	numeric	values in µg/L; if not given by the lab, they are replaced as: • LOD and LOQ known: -1 for $X <$ LOD and -2 for LOD $\leq X <$ LOQ • LOQ known, LOD not: -3 for $X <$ LOQ • LOD known, LOQ not: -1 for $X <$ LOD LOD: limit of detection and LOQ: min. concentration level at which a substance can be measured accurately and reported with certainty.
biomarker_lod	LOD of the biomarker (e.g. mbzp_lod)	numeric	µg/L
biomarker_loq	LOQ of the biomarker (e.g. mbzp_loq)	numeric	µg/L
biomarker_crt	biomarker values standardised for creatinine (e.g. mbzp_crt)	numeric	µg/g crt
biomarker_log	ln-transformed biomarker values	numeric	µg/L
biomarker_crtlog	ln-transformed biomarker values standardised for creatinine	numeric	µg/g crt
biomarker_imp	imputed biomarker values	numeric	µg/L values indicated as: –1 (below LOD), –2 (between LOD and LOQ), or –3 (below LOQ) single random imputation from a truncated lognormal distribution. Imputation performed, if at least 30% of values were detected.
biomarker_impcrt	imputed biomarker values standardised for creatinine	numeric	µg/g crt
biomarker_implog	ln-transformed imputed biomarker value	numeric	µg/L
biomarker_impcrtlog	ln-transformed imputed biomarker values standardised for creatinine (e.g. mbzp_impcrtlog)	numeric	µg/g crt

3 Methodology

To ensure that estimates of biomarker exposure accurately reflect the heterogeneity of the EU-27 children (6-12 years old) population, key socio-demographic and geographic factors, also included as variables in the HBM4EU dataset, were considered as they could influence environmental contaminant levels. In particular, *region* could capture broad environmental and regulatory differences (e.g. variations in industrial activity, climate, and lifestyle) across the European regions. The *degree of urbanisation* (DEGURBA) distinguishes urban, towns/suburbs, and rural settings, which differ in terms of population density, housing characteristics, and potential sources of biomarker exposure. Finally, *household education* (ISCED strata) may serve as a proxy for socioeconomic status, reflecting differences in consumer behavior, dietary patterns, and awareness of chemical risks. Data breakdown is not always available by sex stratum and the age group margins were defined along the range of the HBM4EU dataset: 6-12 years. The external population margins were combined to form a reference grid, which was then used to generate population-weighted estimations in order to derive EU-27 standardized biomarker concentrations.

3.1 Identifying EU-27 population margins

Firstly, age-structured Eurostat data (2023) were used to aggregate population counts per region, accounting for both females and males aged 6–12 yrs (Figure 1).The Eurostat dataset[10] covers all European countries, but only the EU-27 members were retained for the calculations hereby. Populations were then summed by European region, according to the UN geoscheme. In total, ca. 32 M children aged 6–12 years were calculated. The West accounted for 14.08 M, the South for 8.76 M, the East for 6.46 M, and the North for 2.71 M. To note that the sex–stratified Eurostat tables (EU-27, 2023), indicate a nearly even sex split in total, with 6-12 yrs males accounting for 51.4 % of the EU-27 population and and females 48.6 %, respectively. Age-specific counts, for both males and females within each region, appear remarkably uniform across the 6–12 yr span, with each year of age comprising approximately 14% of its regional population (Figure 2). For instance, in the West the proportion ranged from 14.13% at age 6 to 14.67% at age 12; comparable patterns were observed in the North (14.01%–14.95%), East (14.48%–14.94%) and South (13.38%–15.46%).

A contingency table jointly stratifying the population by age (or sex) and DEGURBA could not be identified in the available Eurostat releases. To estimate the distribution of children aged 6–12 years across DEGURBA classes, *urban*, *towns/ suburbs*, and *rural*, two Eurostat datasets were used: (i) population data by EU-27 country (2023), including total population and the population of 6–12-year-old females and males, and (ii) the percentage distribution of the total population per EU-27 country and by DEGURBA class (2020) [10, 11]. For each EU-27 country, the proportion of the population living in each DEGURBA class was multiplied by the total population and by the relative share of children aged 6–12 in that population. This yielded an estimated number of children within each class. The estimates were subsequently aggregated by region following the UN geoscheme. For example, in the West, approximately 3.56 M children were estimated to live in rural areas, 5.01 M in towns and suburbs, and 5.50 M in urban areas. Distributions were also calculated for the remaining regions, enabling cross-regional counts based on level of urbanisation (Figure 3).

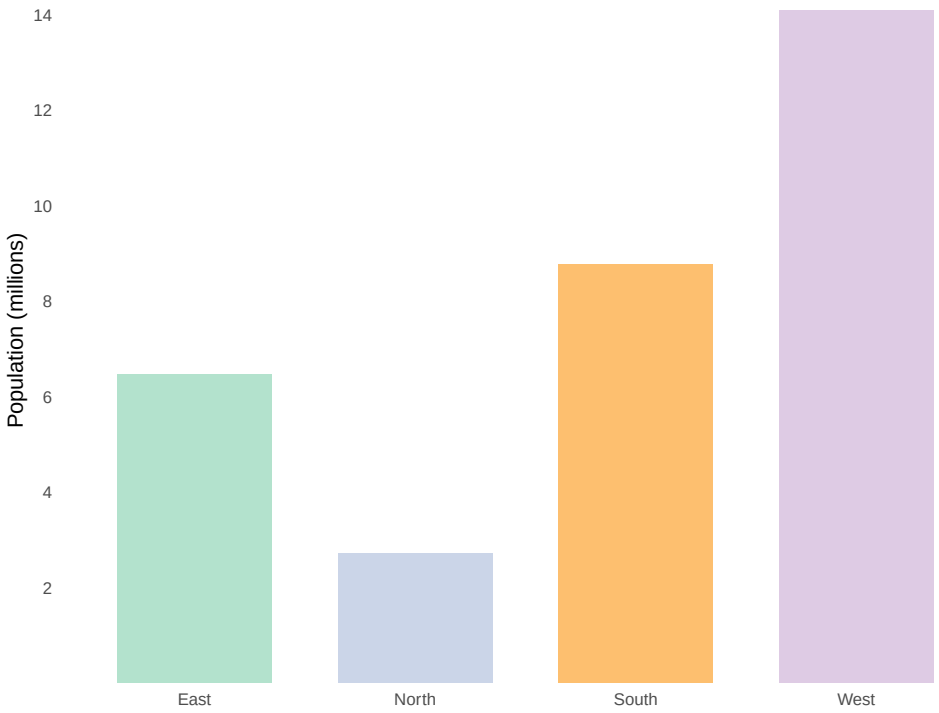


Figure 1: Regional distribution of EU-27 children (6–12 yrs), estimated from Eurostat

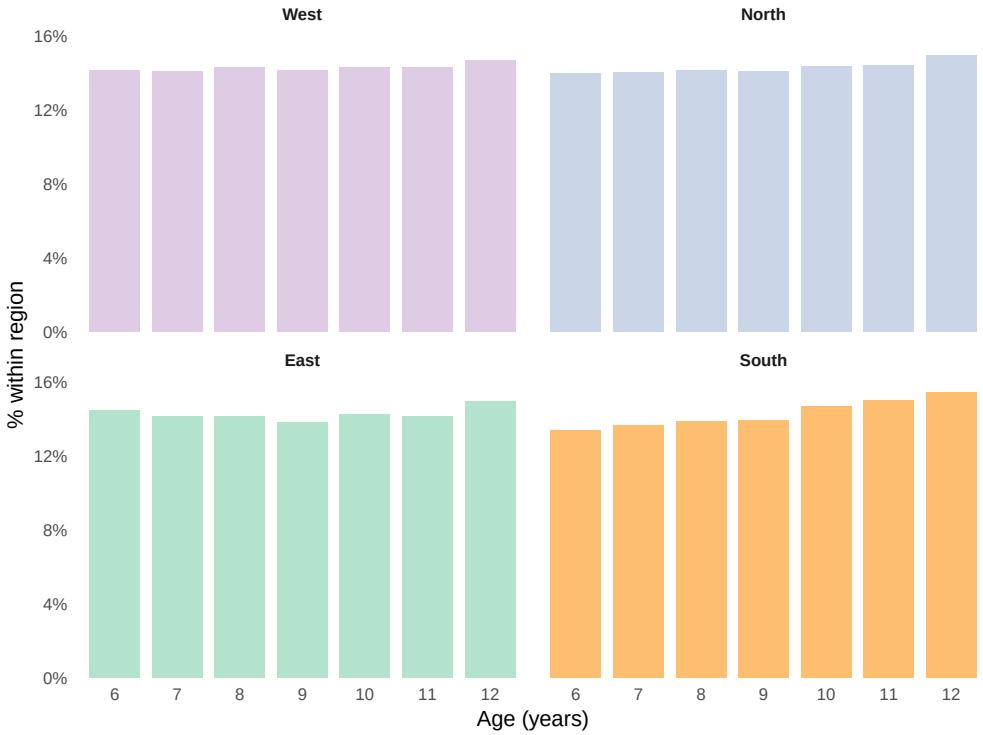


Figure 2: Age distribution of EU-27 children by region, estimated from Eurostat

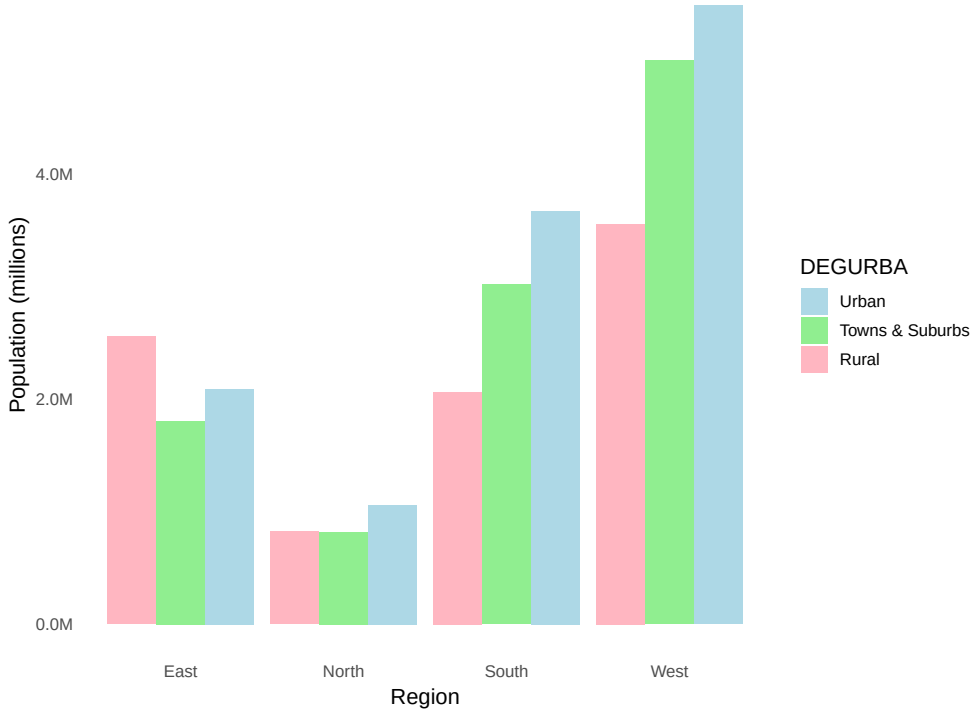


Figure 3: Regional distribution of EU-27 children (6–12 yrs) by DEGURBA, Eurostat estimates

An attempt was also made to estimate the distribution of children across ISCED levels, using Eurostat's "Distribution of households by educational attainment level of the reference person" (2020).[12] Labelled by Eurostat as "experimental", these data may be revised in future releases. The dataset reports the % of households per country in each ISCED category: 0 = early childhood education, 1 = primary education, 2 = lower secondary education, 3 = upper secondary education, 4 = post-secondary non-tertiary education, 5 = short-cycle tertiary education, and 6–8 = tertiary education excluding short-cycle. Since 2020 figures were unavailable for Italy and Sweden, they were substituted with their 2010 and 2015 values, respectively. The ISCED-level shares for all EU-27 countries were then aggregated into 3 strata: Low (ISCED 0–2), Medium (ISCED 3–4) and High (ISCED $\geq$ 5), consistent with the HBM4EU grouping. Each country's Low/Medium/High proportions were then normalized to sum to 100 %. This was done as the original Eurostat dataset provides % for each of the seven ISCED levels, and these values are often subject to rounding and may include small "unknown" or unclassified portions. When summing the ISCED-level percentages for Low, Medium, and High, it can yield a total marginally different from 100%. This step corrects for any residual rounding error and guarantees a proper three-category breakdown for subsequent regional aggregation. The normalized shares were joined with the totals of 6–12 yrs old children and with the UN geoscheme regional mapping. By multiplying each country's total children population by its normalized ISCED-stratum percentages, counts of children in each category were derived. Summing the counts across the UN-defined European regions yielded a regional breakdown (Fig. 4). For example, in the West, children appear most often in medium-education households (ca. 6.08 M), followed by high-education (ca. 5.44 M), and low-education (ca. 2.56 M).

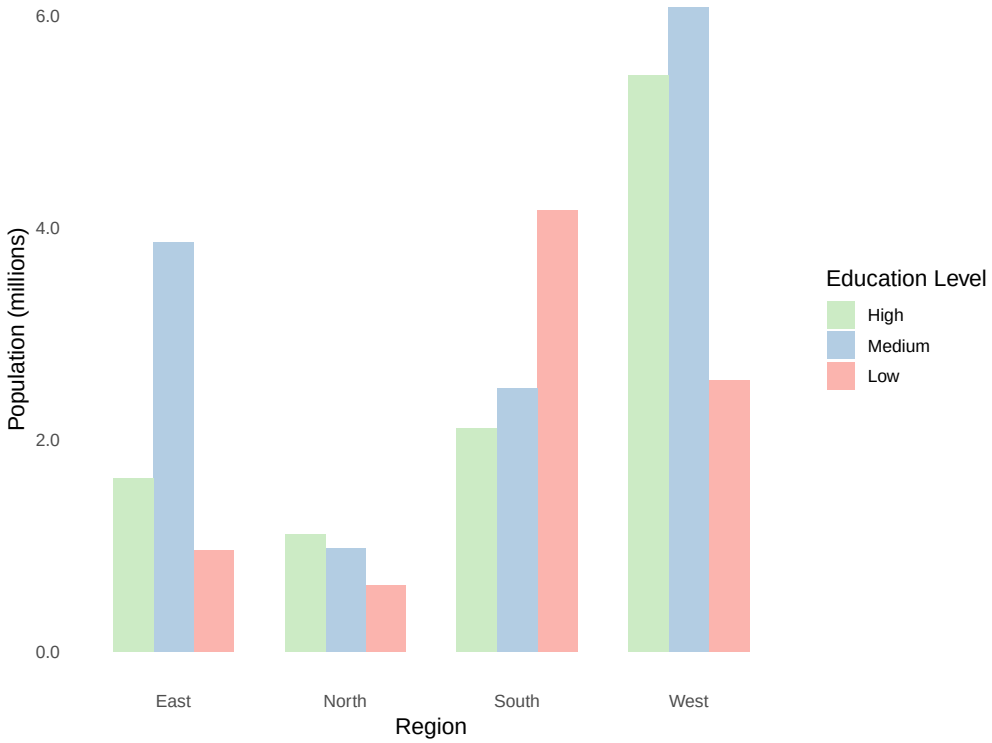


Figure 4: Regional distribution of EU-27 children (6–12 yrs) by ISCED, Eurostat-based estimates

3.2 EU-27 population-standardised reference grid and weights derivation

To obtain an EU-27-standardised mean for children aged 6–12 years, a $4 \times 2 \times 4 \times 3 \times 3 = 288$-cell reference grid was built by crossing external Eurostat margins (Section 3.1) for: *region*, *degree of urbanisation*, *household education* and *sex*; each *season* was given an a-priori weight of 0.25. As detailed in Section 3.1, Eurostat tabulations reveal an essentially uniform age structure within each EU-27 region, with each single-year age from 6 to 12 years accounting for roughly 14 % of the regional child population. Therefore, the seven single–year age bands were collapsed by fixing age at the midpoint (9 years) in the reference grid, and the centered age variable ($\text{age}_c = \text{age} - 9$) can be used in all subsequent regression steps. The 288-cell grid was thereby preserved, and model intercepts correspond to a representative child at the midpoint age of 9-years-old. For any future analyses requiring broader age bands (e.g., adults) or finer spatial resolution, the margins should be re-crossed with single-year (or grouped) ages and the relevant geographic units, thereby expanding the grid in exchange for greater demographic resolution.

Hereby, for every grid cell defined by region ($r \in \{\text{North, South, West, East}\}$), sex ($s \in \{\text{M,F}\}$), season ($q \in \{\text{spring}, \ldots, \text{winter}\}$), degree of urbanisation ($d \in \{\text{urban}, \text{towns/suburbs}, \text{rural}\}$) and household-education stratum ($e \in \{\text{low}, \text{medium}, \text{high}\}$) an external weight was assigned as the product of the marginal Eurostat proportions,

$$w_{rsqde} \;=\; \pi_r^{(\text{reg})}\, \pi_s^{(\text{sex})}\, \pi_q^{(\text{season})}\, \pi_d^{(\text{deg})}\, \pi_e^{(\text{edu})}, \qquad \sum_{r,s,q,d,e} w_{rsqde} = 1.$$

This construction post-stratifies to the EU-27 population under the working assumption that the margins are mutually independent. To the best of our knowledge, most of the current Eurostat releases provide one-way tabulations; hence it might be challenging to obtain the full five-way joint distribution. The independence assumption is thus the least restrictive choice that still yields a complete set of cell weights.

3.3 Model-based, EU-27 population-standardized predictions

3.3.1 Multiple linear regression - no interactions

Multiple Gaussian linear regression (Model 1) was fitted to the ln-transformed, imputed, creatinine-standardised biomarker. This follows the HBM4EU harmonised data-management protocol, in which urinary biomarker concentrations are expressed per gram creatinine to correct for between-sample variation in urine dilution; and values below the analytical LOD/LOQ were stochastically imputed from a truncated log-normal distribution fitted to the observed data above limit [6].

$$Y_i \;=\; \beta_0 \;+\; \sum_{k=1}^{K} \beta_k X_{ik} \;+\; \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, n.$$

Hereby, $Y_i = \text{mbzp\_impcrtlog}_i$ is the natural-log, creatinine-standardised mbzp concentration for individual i. The covariate vector $\mathbf{X}_i$ contains the categorical factors region, DEGURBA, household education, sex, and sampling season, together with the centred age term $\text{age}_\text{c} = \text{age} - 9$. Centering aligns the intercept with the reference-grid profile (Section 3.2) and can reduce intercept–slope collinearity. Categorical predictors were coded as indicator (dummy) variables, taking *North*, *male*, *urban*, *low education* (ISCED 0–2) and *spring* as reference levels. With age centered, the intercept β_0 represents the expected ln-mbzp for a nine-year-old boy in a low-education household, urban setting, living in the North of Europe, and sampled in spring.

The fitted model was evaluated at each of the 288 covariate profiles $g = 1, \ldots, 288$ in the EU-27 population-standardised reference grid, where every profile has $\text{age}_\text{c} = 0$ (age = 9 yrs). This yielded cell-level predictions $\hat{Y}_g$ with model-based standard errors $\text{SE}(\hat{Y}_g)$. Let w_g be the external weight for cell g ($\sum_g w_g = 1$). The EU-standardised mean is:

$$\overline{Y} \;=\; \sum_{g=1}^{288} w_g \, \hat{Y}_g.$$

R's `predict.lm` returns $\hat{Y}_g$ and $\text{Var}(\hat{Y}_g)$ for each design point, but not for a weighted aggregate such as $\overline{Y}$. Post-stratifying the 288 cell predictions with the EU weights overcomes this limitation. Because the model is linear and Gaussian, $\overline{Y}$ is itself linear in the estimated coefficients, and the delta method provides its exact variance:

$$\text{Var}(\overline{Y}) \;=\; \sum_{g=1}^{288} w_g^2 \; \text{Var}(\hat{Y}_g), \qquad \text{SE}_\Delta(\overline{Y}) \;=\; \sqrt{\sum_{g=1}^{288} w_g^2 \, [\text{SE}(\hat{Y}_g)]^2}.$$

Exponentiating $\overline{Y}$ and its $\pm 1.96\,\text{SE}(\overline{Y})$ limits transforms the results to concentration units (µg g^{-1} creatinine), yielding the EU-adjusted geometric mean and its 95% confidence interval on the original scale while preserving the EU-27 demographic composition encoded in the reference grid.

3.3.2 Multiple linear regression - screening and evaluation of interaction terms

Five two-way interactions were examined on the following hypotheses: *region* × *season*, as it may capture climate-driven and behavioral seasonal differences across Europe; *region* × *DEGURBA* due to potential urban–rural exposure contrasts driven by differences in population density, and local pollutant sources and how those may vary across EU; *season* × *DEGURBA* to assess whether seasonal variation may differ by living environment; as well as region × age_c as it could allow region-specific age slopes; and *sex* × age_c for sex-dependent age trends, where age_c is child's age centred at nine years. Each interaction was first added singly to the main-effects model ($\text{AIC}_0 = 8507$); all main effects were retained in every comparison, computing:

$$\Delta\text{AIC} = \text{AIC}_{\text{aug}} - \text{AIC}_0 \quad \text{and} \quad \text{LR}(2\,\text{df}).$$

Retention of interaction terms considered statistical criteria ($\Delta\text{AIC} \leq -2$ and LR $p < 0.05$) and especially biological plausibility. Three-way interactions were omitted to keep the model parsimonious and to maintain stable predictions for every reference-grid cell.

Let $i = 1, \ldots, n$ index study participants and $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})^\top$ contain the K main-effect covariates used in Model 1, including the centered age term $\text{age}_\text{c} = \text{age} - 9$. Suppose a subset $\mathcal{P}$ of predictor pairs (p, q) has been selected for inclusion as two-way interactions (e.g. region×season, region×DEGURBA). For each chosen pair, we define the interaction covariate $W_{ipq} = X_{ip}X_{iq}$ and collect all such terms in the vector $\mathbf{Z}_i = \{W_{ipq} : (p, q) \in \mathcal{P}\}^\top$. The augmented regression fitted in every subsequent analysis is:

$$Y_i = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta} + \mathbf{Z}_i^\top \boldsymbol{\gamma} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $\boldsymbol{\beta}$ contains all main-effect coefficients and $\boldsymbol{\gamma}$ the interaction coefficients. Main effects are retained regardless of which interactions are present; setting any component of $\boldsymbol{\gamma}$ to zero simply removes the corresponding interaction without altering the baseline structure of the model.

For every retained model m, the following were predicted: $\hat{Y}_g^{(m)}$ with standard error $\text{SE}(\hat{Y}_g^{(m)})$ for each grid cell $g = 1, \ldots, 288$. With EU weights w_g ($\sum_g w_g = 1$):

$$\overline{Y}^{(m)} = \sum_g w_g \, \hat{Y}_g^{(m)}, \qquad \text{SE}_\Delta(\overline{Y}^{(m)}) = \sqrt{\sum_g w_g^2 \, \text{SE}(\hat{Y}_g^{(m)})^2}.$$

Exponentiating $\overline{Y}^{(m)}$ and its ± 1.96 $\text{SE}(\overline{Y}^{(m)})$ bounds returns the estimate to the original concentration scale ($\mu\text{g}\,\text{g}^{-1}$ creatinine), yielding the EU-standardised geometric mean and its 95 % confidence interval. Because identical population weights are used in every model, these geometric means remain directly comparable across specifications and preserve the EU-27 demographic composition.

3.3.3 Mixed models - random intercept specification

A random–intercept mixed model was fitted for the natural-log–transformed, creatinine-standardised mbzp concentration. "Cohort" ($j = 1, \ldots, 11$) was modelled with a random intercept $b_j \sim \mathcal{N}(0, \sigma_u^2)$, assumed independent of the residual errors and of all fixed-effect predictors; this term captures unmeasured, study-specific heterogeneity. All parameters were estimated by maximum likelihood, permitting likelihood-ratio tests of nested models. Since the EU-27 population-standardised reference grid fixes age at the midpoint (Section 3.2), age was centered as $\text{age}_c = \text{ageyears} - 9$. Accordingly, the fixed intercept β_0 refers to a nine-year-old child in all reference-category covariate levels. Centering and the retained interaction terms follow the rationale in Section 4.2, making the mixed-model analysis comparable with the fixed-effects models.

$$Y_{ij} \;=\; X_{ij}^{\top}\beta + b_j + \varepsilon_{ij}, \qquad b_j \sim \mathcal{N}(0, \sigma_u^2),\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2),$$

The residual variance σ_e^2 is assumed to be constant (homoscedastic) across cohorts; diagnostic plots (Appendix A) show no meaningful violation. Here, $i = 1, \ldots, n_j$ index individuals within cohort j and $j = 1, \ldots, 11$ index the cohorts of the HBM4EU dataset related to children. The design vector X_{ij} contains: region (4 levels; North ref., 3 dummies), sex (M = 0, F = 1), centered age_c, DEGURBA (urban ref., 2 dummies), household education (ISCED 0–2 ref., 2 dummies), and sampling season (spring ref., 3 dummies). Introducing two-way interactions, let $\mathbf{R}_i \in \{0,1\}^3$ (region), $\mathbf{S}_i \in \{0,1\}^3$ (season), and $\mathbf{D}_i \in \{0,1\}^2$ (DEGURBA). With the two retained blocks (region×season, region×DEGURBA) the model becomes:

$$Y_{ij} = X_{ij}^{\top}\beta + (\mathbf{R}_i \otimes \mathbf{S}_i)^{\top}\theta + (\mathbf{R}_i \otimes \mathbf{D}_i)^{\top}\gamma + b_j + \varepsilon_{ij},$$

where $\theta \in \mathbb{R}^9$ and $\gamma \in \mathbb{R}^6$ parameterise the region–season and region–DEGURBA contrasts. $\mathbf{R}_i \otimes \mathbf{S}_i$ denotes the element-wise product of the three-level region dummy $\mathbf{R}_i$ and the three-level season dummy $\mathbf{S}_i$, yielding the nine region–season contrasts (and similarly for $\mathbf{R}_i \otimes \mathbf{D}_i$).
For both mixed models, the intraclass-correlation coefficient is defined as:

$$\widehat{\text{ICC}} \;=\; \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2},$$

where $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ are the variance estimates of the random intercept and the residual error.

Each mixed model was projected onto the EU-27 population-standardised reference grid. Because the grid contains only the midpoint age, the centered age regressor does not enter the linear predictors used for EU standardisation, yet its coefficient β_{age} and its sampling variance are still estimated before from the complete 6–12 years HBM4EU sample. Let $\hat{\eta}_g = x_g^{\top}\hat{\beta}$ be the fixed-effects predictor for grid cell g. The EU-standardised log-mean is:

$$\hat{\mu} \;=\; \sum_{g=1}^{288} w_g\, \hat{\eta}_g,$$

Three methods were used to quantify $\text{SE}(\hat{\mu})$. Each method was applied with the same 288-cell reference grid and external weights across all model specifications, so the resulting standard errors are directly comparable:

a) Delta method: Linearising the cell–specific predictor $\hat{\eta}_g = x_g^\top \hat{\beta}$ yields:

$$\mathrm{Var}(\hat{\eta}_g) \;=\; x_g^\top \, \widehat{\mathrm{Var}}_{\mathrm{ML}}(\hat{\beta}) \, x_g.$$

Adding the estimated random–intercept variance $\hat{\sigma}_u^2$ to every cell and aggregating with the external weights $\{w_g\}_{g=1}^{288}$ gives:

$$\mathrm{SE}_\Delta \;=\; \sqrt{\sum_{g=1}^{288} w_g^2 \Big[x_g^\top \, \widehat{\mathrm{Var}}_{\mathrm{ML}}(\hat{\beta}) \, x_g + \hat{\sigma}_u^2 \Big]}.$$

*Note:*Both the Delta-method and the following Monte-Carlo schemes treat the random-intercept variance estimate $\hat{\sigma}_u^2$ as *fixed*, i.e. they ignore the small sampling variability in $\hat{\sigma}_u^2$ itself. Given the large sample size, treating $\hat{\sigma}_u^2$ as fixed could be an acceptable simplification; accounting for its sampling error is expected to widen the CIs marginally.

The closed-form expression for SE_Δ: (i) propagates fixed-effect uncertainty through $\widehat{\mathrm{Var}}_{\mathrm{ML}}(\hat{\beta})$ and (ii) incorporates between-cohort heterogeneity via $\hat{\sigma}_u^2$. Under large-sample theory $\hat{\beta}$ is approximately multivariate normal, and $\hat{\sigma}_u^2$ is treated as independent of $\widehat{\mathrm{Var}}_{\mathrm{ML}}(\hat{\beta})$. A Wald 95 % interval on the concentration scale is: $[\exp(\hat{\mu} \pm 1.96\,\mathrm{SE}_\Delta)]$.

b) Monte-Carlo sampling of β ("MC–fixed"). To propagate uncertainty from the fixed effects only, $M = 5{,}000$ draws were taken from the asymptotic sampling distribution of the maximum-likelihood estimator:

$$\beta^{(m)} \sim \mathcal{N}\big(\hat{\beta},\, \widehat{\mathrm{Var}}_{\mathrm{ML}}(\hat{\beta})\big), \qquad m = 1, \ldots, M.$$

For each draw m and every reference-grid cell g the linear predictor is:

$$\eta_g^{(m)} = x_g^\top \beta^{(m)}$$

The EU-weighted mean in replicate m and its Monte-Carlo standard error are:

$$\mu_{\mathrm{fix}}^{(m)} = \sum_{g=1}^{288} w_g \, \eta_g^{(m)}, \qquad \mathrm{SE}_{\mathrm{fix}} = \mathrm{sd}\big\{\mu_{\mathrm{fix}}^{(m)}\big\}_{m=1}^{M}$$

Because $\hat{\sigma}_u^2$ is not resampled here, $\mathrm{SE}_{\mathrm{fix}}$ reflects fixed-effect uncertainty alone; additional variation from between-cohort heterogeneity is incorporated in the subsequent "MC–full" procedure.

c) Monte-Carlo sampling of both β and the random intercept ("MC–full"). To propagate fixed-effects *and* between-cohort uncertainty, each of the $M = 5{,}000$ replicates proceeds in two steps. *Step 1:* drawing the fixed-effect vector:

$$\beta^{(m)} \;\sim\; \mathcal{N}\big(\hat{\beta}, \widehat{\mathrm{Var}}_{\mathrm{ML}}(\hat{\beta})\big), \qquad m = 1, \ldots, M.$$

Step 2: drawing a single cohort-level intercept, common to all grid cells in that replicate:

$$u^{(m)} \;\sim\; \mathcal{N}\big(0, \hat{\sigma}_u^2\big).$$

For every grid cell $g = 1, \ldots, 288$ the linear predictor is:

$$\eta_g^{(m)} = x_g^\top \beta^{(m)} + u^{(m)}$$

The EU-weighted mean in replicate m and its Monte-Carlo standard error are:

$$\mu_{\text{full}}^{(m)} = \sum_{g=1}^{288} w_g \, \eta_g^{(m)}, \qquad \text{SE}_{\text{full}} = \text{sd}\{\mu_{\text{full}}^{(m)}\}_{m=1}^{M}$$

With $M = 5{,}000$ draws, the Monte-Carlo sampling error in SE_{full} is < 0.01.

Comparing SE_{Δ} with SE_{fix} assesses the delta-method approximation, whereas the difference between SE_{fix} and SE_{full} quantifies the additional uncertainty introduced by cohort heterogeneity. For every model the point estimate $\exp(\hat{\mu})$ is reported as the EU-standardised geometric mean, accompanied by the concentration-scale 95 % confidence interval derived from the chosen standard-error method (Delta, MC–fixed, or MC–full).

3.4 Non-model, design-based methods

3.4.1 Direct post-stratification weighting

An unweighted mean of the log-transformed, creatinine-standardised mbzp was first computed on the full eligible sample of children aged 6–12 years, excluding 39 records with missing biomarker values (resulting to n = 2 784). Moroever, it was calculated for the dataset which was then further restricted to records with complete information on the five post-stratification variables—EU-27 region, sex, sampling season, DEGURBA, and household ISCED—yielding the post-stratification sample (n = 2 722; 63 records excluded: 62 missing ISCED and 1 missing DEGURBA). Direct post-stratification weights were assigned by matching each of the 2 722 observations to the Eurostat reference grid of joint marginal proportions for region, sex, season, DEGURBA, and ISCED (Section 3.2). Note that the sample size is equivalent to the one in Section 3.3, considering that by default both lm() and lmer() omit any observations with missing values on the outcome or predictors.

Let $h = 1, \ldots, H$ index the H strata defined by those factors, and N_h be the number of sampled children in stratum h. If M_h is the Eurostat population share for stratum h ($\sum_h M_h = 1$), each child i in stratum h received a raw weight:

$$w_i^{\text{raw}} \;=\; \frac{M_h}{N_h}, \quad i = 1, \ldots, N_h.$$

Because excluding the 62 post-stratification-incomplete records removed approximately 14.6 % of the total raw weight mass, the remaining weights were renormalised to sum to unity:

$$w_i \;\leftarrow\; \frac{w_i^{\text{raw}}}{\sum_{k=1}^{N} w_k^{\text{raw}}}, \qquad N = 2\,722$$

The EU-standardised mean of log-transformed, creatinine-standardised mbzp (`mbzp_impcrtlog`) was then computed as:

$$\overline{y}_{\text{weighted}} \;=\; \sum_{i=1}^{N} w_i \, y_i,$$

where y_i denotes child i's log-mbzp concentration (standardised for creatinine). *Sensitivity analysis:* missing ISCED values (n = 62) were imputed to the "Medium" category while retaining all other

cases; post-stratification weights and the weighted mean were recomputed to confirm that excluding incomplete records had a negligible effect.

Overall, this weighted-mean procedure does not yield a closed-form variance; standard errors must be derived with design-based methods (e.g. Taylor linearisation using the R `survey` package) or with non-parametric bootstrap resampling.

3.4.2 Survey-design weighted estimation

The **survey** package [16] was used to estimate the EU-standardised geometric mean of log-transformed mbzp (creatinine adjusted). Post-stratified weights were attached to every child and supplied to `svydesign` under two sampling specifications:

- **Independent design** (`ids = ~1`): each child is treated as a primary sampling unit (PSU); the SE therefore reflects the variance inflation from unequal weights.
- **Clustered design** (`ids = ~cohort`): each HBM4EU cohort is treated as a PSU, allowing the SE to incorporate both weight variability and any within-cohort correlation. Comparing the two CIs quantifies any precision loss due to clustering.

For either design, `svymean` applies *Taylor-series linearisation*, expanding the weighted mean $\overline{y}_w = \sum_i w_i y_i$ to first order around its expectation and providing a design-consistent SE without resampling. The log-scale estimate $\hat{\mu}$ and SE were back-transformed to obtain:

$$\widehat{\mu}_{\mathrm{s}} = \exp(\hat{\mu}), \qquad 95\ \%\ \mathrm{CI} = \exp\!\big(\hat{\mu} \pm 1.96\,\widehat{\mathrm{SE}}\big).$$

Weight heterogeneity was summarised by the design effect $\mathrm{DEFF} = n \sum_i w_i^2$, yielding an effective sample size $n_{\mathrm{eff}} = n/\mathrm{DEFF}$. Range, quartiles and region-specific DEFFs were inspected to inspect for outlying weights or high-variance strata. Robustness to extreme weights was assessed by trimming at the 99% and 95% percentiles, renormalising to $\sum_i w_i = 1$, and re-estimating the geometric mean; checking whether trimmed estimates stayed within the untrimmed 95 % CI, confirming whether there was limited influence of the largest weights.

Sensitivity margin for single-year age: Considering that Eurostat data showed that 6–12-year-olds (males and females) are distributed almost uniformly across single ages within every EU-27 region (Section 3.1), a diagnostic calibration was run, adding a sixth margin *region × age*. The existing five-way post-stratification *frequency* weights were raked to a synthetic table in which each region's total weight was split equally across the seven *observed* ages (6–12 y), restricting the table to age–region combinations present in the HBM4EU sample to avoid inflating weights for unsampled strata. Regional and overall totals therefore remained unchanged. The survey design still treated cohorts as primary sampling units (PSUs) and introduced no additional covariates; the aim was to gauge whether age imbalance could bias the EU-level estimate and how it would affect the weights-only design effect (DEFF) and the 95 % confidence interval (CI). The calibration was repeated without cohort-clustering as well.

3.4.3 Raking calibration

An iterative proportional adjustment procedure ('raking') was applied to $n = 2\,722$ HBM4EU children with complete biomarker and auxiliary data. While direct post-stratification forces the sample to match the full Eurostat cross-classification of region, sex, sampling season, DEGURBA and household ISCED; raking instead calibrates each marginal distribution separately. This "softer" constraint smooths the most extreme weights and typically lowers the weights-only design effect and therefore increases the effective sample size.

Beginning the analysis with an unclustered survey design (`ids = ~1`), the initial weights were set to $\omega_i^{(0)} = 1$ for every child $i = 1, \ldots, n$. For each margin $j \in \{\text{region}, \text{sex}, \text{season}, \text{DEGURBA}, \text{ISCED}\}$ and category k (e.g. North vs. South for region, M vs. F for sex), let M_{jk} denote the corresponding Eurostat population proportion ($\sum_k M_{jk} = 1$). These proportions were scaled to the sample size, $T_{jk} = n\, M_{jk}$. At iteration t, the weight of each child in cell (j, k) was updated by

$$\omega_i^{(t+1)} = \omega_i^{(t)} \frac{T_{jk}}{\sum_{r:\, x_{rj}=k} \omega_r^{(t)}}, \qquad x_{ij} = k,$$

where the denominator is the current weighted total in that margin. Iterations cycled over all (j, k) pairs until every weighted marginal matched its target to within the default tolerance $\epsilon = 10^{-6}$. After calibration, *cohort* clustering was re-introduced by re-building the design with `ids = ~cohort` and the calibrated frequency weights $\sum_i \omega_i = n$.

The raked log-mean of creatinine-standardised mbzp was:

$$\hat{\mu}_{\text{rake}} \;=\; \sum_{i=1}^{n} \omega_i\, y_i, \qquad y_i = \log\!\big(\text{mbzp}_i\big),$$

where ω_i are the calibrated (frequency) weights. Its SE was obtained by Taylor linearisation using `svymean` (see also Section 3.4.2). Exponentiating $\hat{\mu}_{\text{rake}}$ and its ± 1.96 SE limits yielded the geometric mean and its 95 % CI. Re-specifying the design with `ids = ~cohort` provided a cluster-robust CI.

For comparability with the post-stratification analysis, the calibrated frequency weights were also rescaled to the probability scale, $\tilde{w}_i = \omega_i / n$ ($\sum_i \tilde{w}_i = 1$). These probability weights were used *only* for diagnostics (design effect, histograms); all point estimates and CIs continued to rely on the frequency weights ω_i. Weight heterogeneity was summarised by the "weights-only" design effect:

$$\text{DEFF} = n \sum_{i=1}^{n} \tilde{w}_i^{\,2} = \frac{1}{n} \sum_{i=1}^{n} \omega_i^2, \qquad n_{\text{eff}} = \frac{n}{\text{DEFF}},$$

and by the empirical range and quartiles of the calibrated weights.

4 Results & Discussion

4.1 Exploratory data analysis

The HBM4EU sample (n = 2 784; 39 missing values for mbzp_impcrtlog) included the following regional breakdown: 21.6% children from North, 21.9% from South, 25.8% from West, and 30.8% from East (Figure 5). By contrast, population shares for 6–12-year-olds (see section 3.2; Figure 1), based on EU-27 Eurostat data, corresponded to: North 8.5%, South 27.4%, West 44.0%, and East 20.2%. To note that the HBM4EU dataset included Norway also, a non-EU-27 country. Overall, it could be deduced that Northern and Eastern Europe are overrepresented, while Western and Southern Europe are underrepresented within the HBM4EU sample for children.

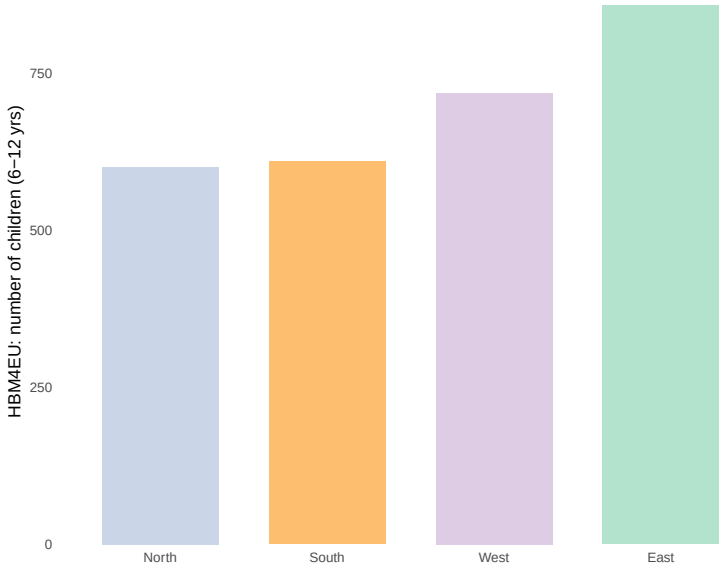


Figure 5: Regional distribution - HBM4EU sample: children aged 6–12 yrs per European region.

Age distribution appeared to vary by region within the HBM4EU dataset: in North and South, more than half of the subjects were of age 7 (51.3% and 55.3%, respectively) with very few 6-years-old (<6%) children. West was more evenly spread between ages 6-12 (28.4% at age 7; 5.2% at age 12), while the East peaked at ages 11 (27.7%) and 9 (24.4%) [Figure 6]. A scatterplot of the standardised log-biomarker concentration versus age, with a LOESS smoother, is illustrated in Figure 7. A linear regression suggests that there is a trend (β = –0.063 log-units/yr, SE = 0.014, p < 0.001%), but the model explains only a small fraction of variance (R^2 = 0.0075, < 1%), underscoring that age alone is a weak predictor. Adding a quadratic term did not improve fit (ΔRSS = 0.053, p = 0.85), supporting linearity.

Figure 8 illustrates that within the HBM4EU sample, subjects residing in urban residence ranged from 18.8% in the West to 64.3% in the North, with South having the highest rural share (34.6%) and West the largest towns/suburbs proportion (48.3%). In contrast, Eurostat EU-27 margins (Fig. 3) appear more balanced: urban shares span only 32.4% (East) to 42.0% (South), with rural and suburban classes each accounting for roughly one-third of children in every region.

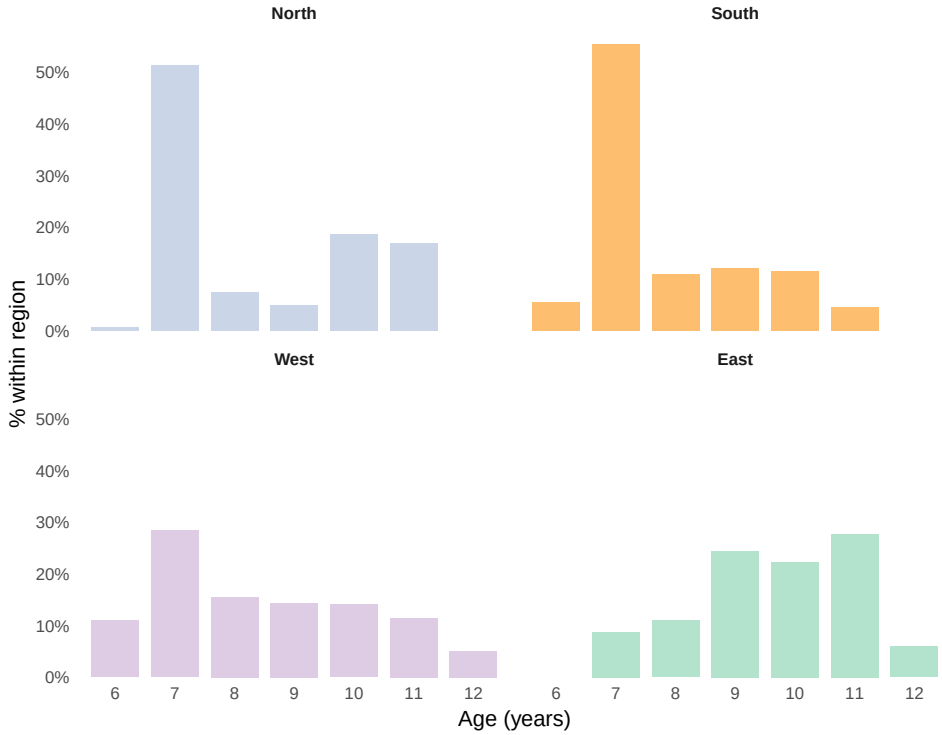


Figure 6: Age distribution - HBM4EU sample: % of children aged 6–12 yrs per European region.

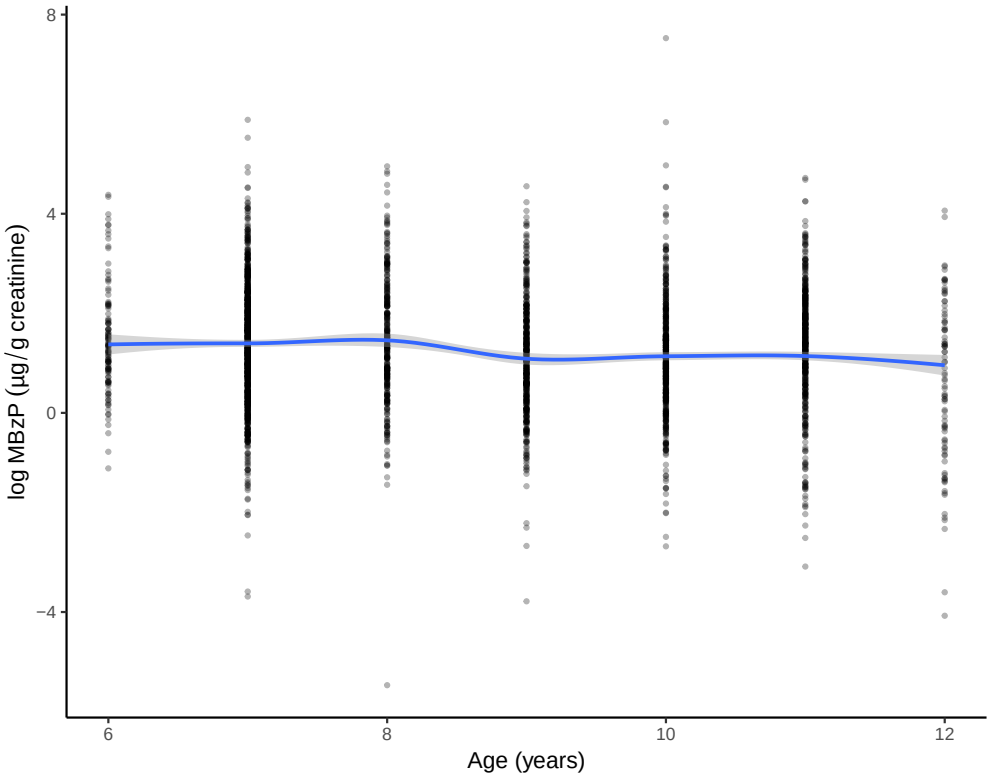


Figure 7: log-mbzp (µg/g crt) vs. age with LOESS smoother - HBM4EU data (children).

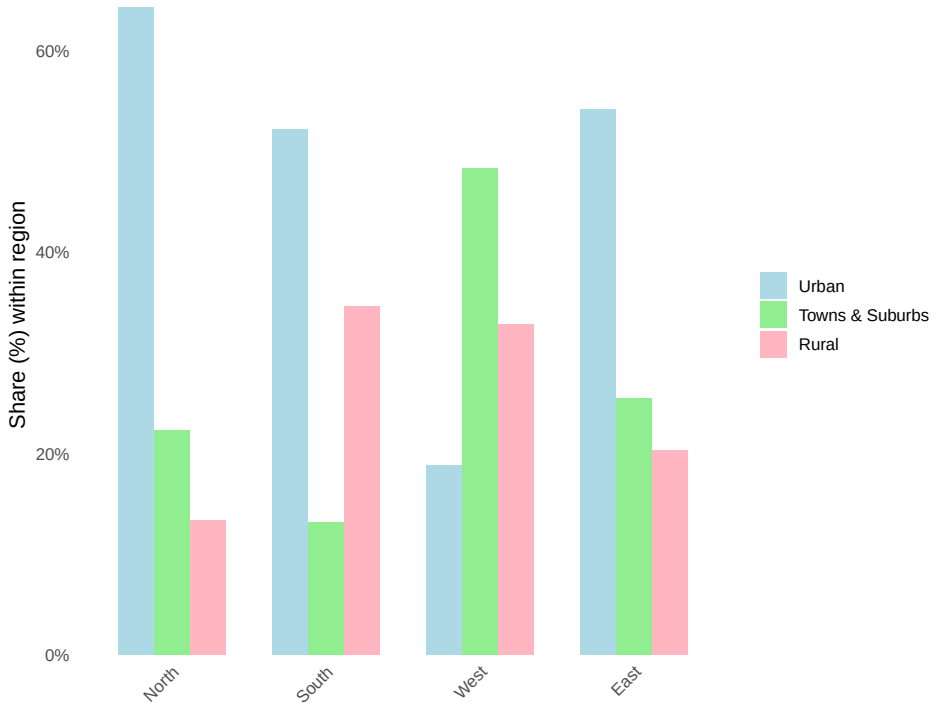


Figure 8: DEGURBA distribution by region - HBM4EU sample: share of children living in Urban, Towns & Suburbs, and Rural settings per European region.

Eurostat EU-27 data (Fig. 4) has suggested that low-ISCED (0-2) households comprise 14.9 %–47.6 % of children (East 14.9 %, South 47.6 %), medium (3-4) 28.3 %–59.8 % (South 28.3 %, East 59.8 %), and high (≥ 5) 24.1 %–40.8 % (South 24.1 %, North 40.8 %). In contrast, the HBM4EU sample (Fig. 9) seems to over-represent high-ISCED households in all regions (North 64.4 %, South 61 %, West 62.1 %, East 50.1 %) and under-represents low-ISCED households (< 7 % vs 14.9 %–47.6 %).

Exploratory analysis of the log-transformed, crt-standardised mbzp concentrations revealed heterogeneity across strata. By region (Fig. 15a; Appendix), median exposures declined from West (1.66; IQR 0.95–2.50) and South (1.49; 0.90–2.16) to North (1.07; 0.39–1.81) and East (0.79; –0.06–1.62), with outliers in West (max 5.84) and North (max 7.53). A gradient appeared across DEGURBA (Fig. 15b; Appendix): median rose from urban areas (1.01; 0.29–1.85) through towns (1.36; 0.71–1.99) to rural settings (1.46; 0.82–2.29). When stratified by ISCED (Fig. 15c; Appendix), children in low-ISCED households had the highest median levels (1.55; 0.74–2.17) and greatest spread, compared with medium (1.26; 0.41–2.12) and high (1.24; 0.57–1.98) ISCED. Furthermore, seasoN boxplots (Fig. 15d; Appendix) showed peak concentrations when sampling was done in summer (1.64; 0.70–2.43) and the lowest in autumn (1.12; 0.38–1.88), while males and females (Fig. 15e; Appendix) exhibited similar distributions (M: 1.26; 0.55–2.04 vs F: 1.25; 0.46–2.02). Overall, these unadjusted patterns highlight between-stratum differences and could also post-stratification weights and multivariable models to obtain unbiased, precision-adjusted estimates.

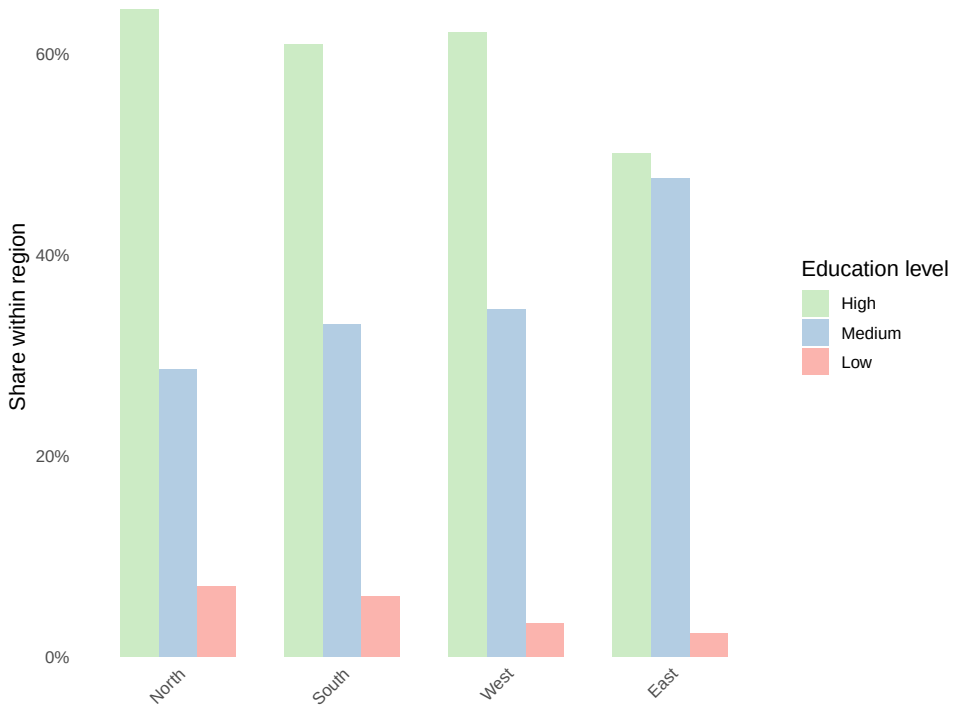


Figure 9: ISCED distribution – HBM4EU sample: share of children in Low, Medium, and High ISCED households across European regions.

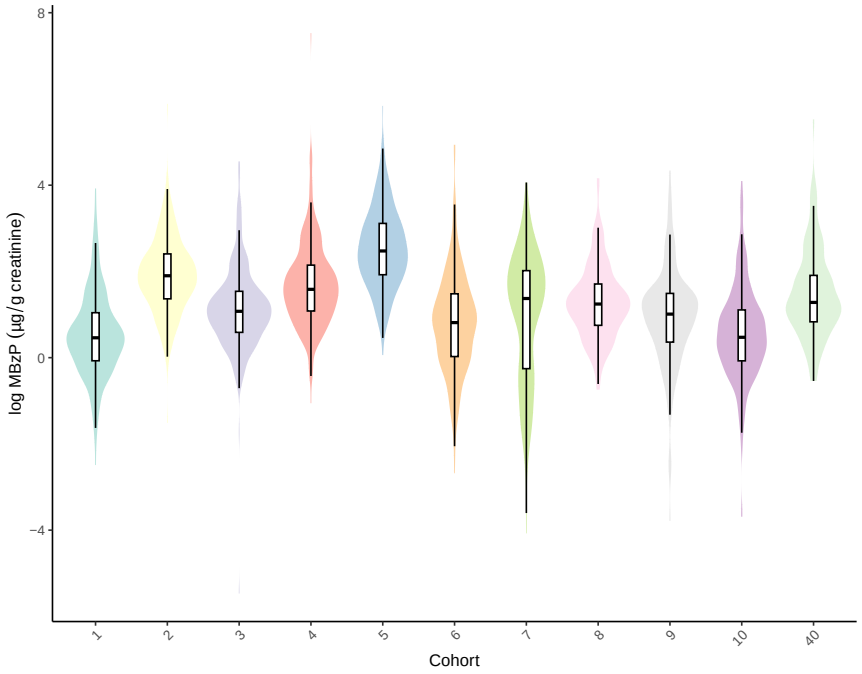


Figure 10: Cohort-level distributions of log-mbzp (µg/g creatinine) -HMB4EU (children). Each "violin" depicts density within a cohort, with an overlaid boxplot showing median and IQR.

Cohort-level log-mbzp (crt standardised) distributions (Figure 10) varied with median values ranging from 0.46 (–0.07–1.04; Cohort 1: C NPHI-InAirQ ; HU) to 2.47 (1.92–3.11; Cohort 5: C-ANSP-ESTEBAN; FR), with Cohort 2 (C-EPIUD-NAC-II; IT) also elevated (1.90; 1.36–2.41). Several cohorts (e.g. 4, 5) exhibited heavy upper tails (max > 5.8), highlighting marked between-site heterogeneity and potentially supporting later the inclusion of a random-intercept term.

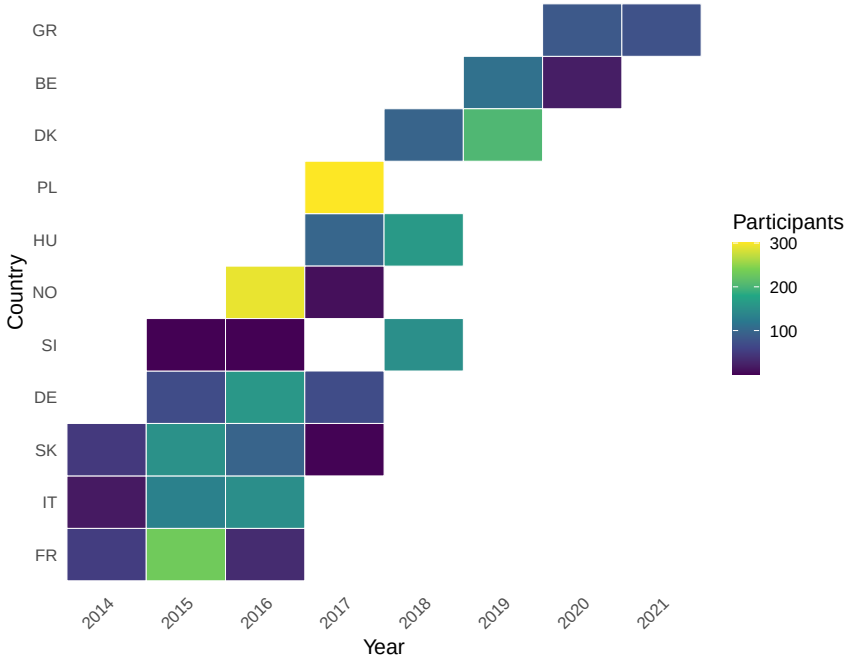


Figure 11: Sampling intensity by country and year (HBM4EU, children-related cohorts). Countries correspond to study cohorts (see Table 2) and are ordered by their first sampling year.

Figure 11 shows a heat-map of sampling intensity (number of participants) for each cohort (country) and by calendar year. Countries are ordered by the first year in which they appear in the HBM4EU children's dataset, and the shading of each tile corresponds to the count of subjects sampled in that country–year. From the plot, five cohorts-France (C-ANSP-ESTEBAN), Italy (C-EPIUD-NAC-II), Slovakia (C-SZU-PCB), Germany (C-UBA-GerES-V) and Slovenia (C-JSI-SLO-CRP)- span three or more sampling years; most of the remaining cohorts cover two years each, and only Poland (C-NIOM-POLAES) appears sampled in a single year. Sampling year and cohort membership may be confounded. One might consider collapsing cohorts with overlapping sampling-year windows into broader time-band groups. The overlaps are though often asymmetric, so some year bands may include some cohorts only partially and we may still leave year bands nearly collinear with the grouped cohort term. Sampling year was excluded from both OLS and mixed-effects models, aiming to allow for the cohort term in the latter to capture study-specific temporal and contextual differences.

Figure 12 shows the $\log_{10}$-transformed weights w_h for each of the 288 reference-grid strata (Section 3.2). On the original scale, w_h ranges from 7.5×10^{-4} to 9.13×10^{-3} (median 3.12×10^{-3};

mean = $1/288 \approx 3.47 \times 10^{-3}$), and no stratum exceeds 1% of the total weight. The ten largest weights all correspond to 9-year-old children in the West region from medium-educated households (urban: 9.13×10^{-3}; towns & suburbs: $\approx 7.89 \times 10^{-3}$). This modest spread in the joint-margin grid confirms that no single stratum drives the EU-standardised mean, thereby suggesting robustness of the weighted-aggregation approach.

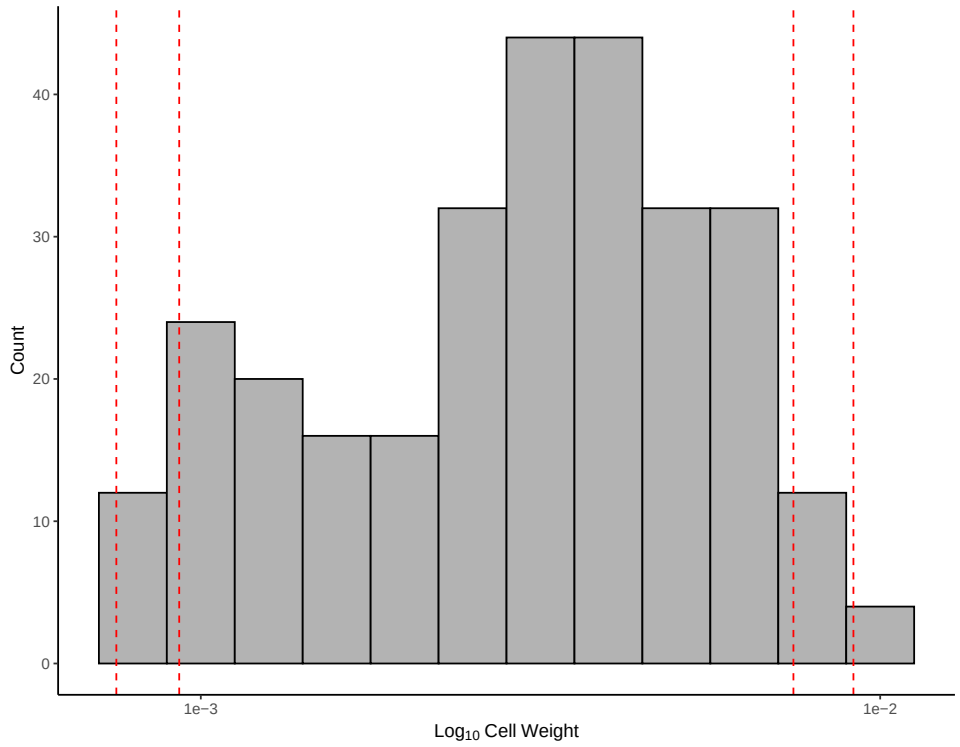


Figure 12: Distribution of the EU27 "reference-grid" of cell weights (n = 288); Section 3.2. Histogram of $\log_{10}$(cell weight); dashed lines at the 1st, 5th, 95th, and 99th percentiles.

4.2 OLS model-based standardization

4.2.1 Multiple linear regression—no interactions; EU-27 standardised values

A Gaussian linear model was fitted to the log–creatinine–standardised outcome $Y_i = \text{mbzp\_impcrtlog}_i$, with predictors region (reference = North), sex (reference = male), centered age ($\text{age}_c = \text{age} - 9$), DEGURBA (reference = urban), household education (reference = low), and sampling season (reference = spring). The model explained a modest proportion of variance ($R^2 = 0.116$, adjusted $R^2 = 0.112$); residual SE = 1.15; $F_{12,2709} = 29.7$, $p < 2 \times 10^{-16}$.

Region: Relative to the North, the regression coefficient for children living in the South was 0.416 log-units (95 % CI 0.277–0.555; $p = 4.9 \times 10^{-9}$). For the West the coefficient was 0.601 log-units (0.467–0.736; $p < 2 \times 10^{-16}$), whereas for the East it was −0.307 log-units (–0.440 to –0.174; $p = 6.5 \times 10^{-6}$). **Degree of urbanisation:** Rural residence was associated with a 0.262 log-unit increase compared with urban areas (0.139–0.385; $p = 3.1 \times 10^{-5}$); the coefficient for"Towns & suburbs" was not statistically different from zero (0.083; –0.032–0.197; $p = 0.16$). **Season:** Winter samples had a coefficient of −0.199 log-units relative to spring (–0.326 to –0.071; $p = 0.0022$); summer showed a weak positive estimate (0.130; –0.019–0.280; $p = 0.087$); autumn did not differ from spring (−0.028; –0.152–0.096; $p = 0.66$). No statistically significant associations were observed for sex, centered age, or household education (all $p > 0.35$).

Table 3: Significant predictors ($p < 0.05$) of log-mbzp in HMB4EU children from an ordinary-least-squares (OLS) linear model.

Effect	β	95 % CI	p
South (vs. North)	0.416	0.277–0.555	4.9×10^{-9}
West (vs. North)	0.601	0.467–0.736	$< 2 \times 10^{-16}$
East (vs. North)	−0.307	−0.440–−0.174	6.5×10^{-6}
Rural (vs. Urban)	0.262	0.139–0.385	3.1×10^{-5}
Winter (vs. Spring)	−0.199	−0.326–−0.071	0.0022

Projecting the model predictions onto the 288-cell EU reference grid (see Section 3.2) and aggregating with EU population weights, yielded a mean log-mbzp of:

$$\overline{Y}_1 = 1.459 \quad (\text{SE}_\Delta = 0.0063),$$

Corresponding to a population-standardised geometric mean of:

$$\exp(\overline{Y}_1) = 4.30\ \mu\text{g g}^{-1}\ \text{creatinine}\ (95\%\ \text{CI}:\ 4.25\text{–}4.36)$$

All uncertainty was obtained analytically via the delta method (see section 3.3.1).

4.2.2 Multiple linear regression - with interactions; EU-27 standardised values

Table 4: Diagnostics for tested OLS models (HBM4EU children's data)

Model	AIC	ΔAIC	LR χ^2 (df)	p-value
main-effects model (no interactions)	8507	0	–	–
region×season	8480	−27	44.1 (9)	1.4×10^{-6}
region×DEGURBA	8383	−124	135.7 (6)	$< 2 \times 10^{-16}$
region×season + region×DEGURBA	8359	−148	177.4 (15)	$< 2 \times 10^{-16}$
DEGURBA×season	8505	−2	13.9 (6)	0.031
region×ageyears	8364	−143	148.3 (3)	$< 2 \times 10^{-16}$
sex×ageyears	8509	+2	0.01 (1)	0.92

Adding the *region* × *season* term lowered AIC (ΔAIC = −27) and yielded a significant LR test ($\chi^2_9 = 44.1$, $p = 1.4 \times 10^{-6}$). Winter–spring contrasts in log(mbzp) were +0.09 in the West (p = 0.44), −0.03 in the North (p = 0.85), while −0.46 in the East (95 % CI [−0.73, −0.19]; $p = 0.001$), and −0.53 in the South (95 % CI [−0.83, −0.24]; $p = 4.6 \times 10^{-4}$. Regional seasonality is consistent with how indoor temperatures and ventilation could modulate BBzP emissions, e.g. from PVC–containing materials. In the West, prolonged winter heating in confined spaces could raise indoor levels, whereas in warmer climates and intermittent ventilation may favor higher summer emissions. Keeping the *region* × *season* term could represent climate-driven exposure shifts. As the underlying emission processes are not age-–specific, the interaction could be applicable across all age-groups.

The addition of *region* × *DEGURBA* term improved the model fit (ΔAIC = −124; $\chi^2_{(6)} = 135.7$, $p < 2 \times 10^{-16}$). Estimated rural-urban contrasts in log(mbzp) differed by region: North, +1.30 (95 % CI [1.01, 1.59]; $p = 2.1 \times 10^{-18}$); East, +0.39 (95 % CI [0.17, 0.62]; $p = 6.1 \times 10^{-4}$); West, +0.15 ($p = 0.23$, NS); and South, −0.37 (95 % CI [−0.58, −0.17]; $p = 4.0 \times 10^{-4}$). Thus, in northern and eastern Europe, rural settings exhibited higher mbzp, whereas in southern Europe the urban excess predominated. These region-specific rural–urban disparities may reflect geographic variation in housing characteristics, local emission sources, or lifestyle factors, justifying the inclusion of the *region*×*DEGURBA* interaction. Since the assumption is that emissions stem from physical/chemical processes (e.g. leaching from building materials) rather than age-dependent physiology, this interaction is expected to be applicable across all age groups. Whether such region × DEGURBA disparities generalize to other exposure biomarkers could be further investigated.

Inclusion of the *DEGURBA*×*season* term yielded a modest improvement in fit (ΔAIC = −2; LR $\chi^2_6 = 13.9$, $p = 0.031$). The rural–urban contrast in log(mbzp) varied by season—largest in summer at +0.64 (95 % CI 0.31–0.97; $p = 1.3 \times 10^{-4}$), intermediate in spring at +0.43 (0.18–0.68; $p = 7.1 \times 10^{-4}$) and winter at +0.20 (0.01–0.39; $p = 0.037$), and small, non-significant in autumn at +0.10 ($p = 0.33$). Overlapping CIs for winter and autumn suggest that any rural–urban difference is minor during cooler seasons; the modest winter excess may reflect heating-related indoor sources

rather than ventilation-driven factors.

The *region*×age_c interaction improved model fit (ΔAIC = −143; $\chi^2_3 = 148.3$, $p < 2 \times 10^{-16}$). Estimated age slopes ($\Delta \log[\text{mbzp}]\,\text{yr}^{-1}$) were +0.305 (95 % CI [0.243, 0.366]; $p = 5.0 \times 10^{-22}$) in the North; –0.215 (95 % CI [–0.286, –0.145]; $p = 2.4 \times 10^{-9}$) in the South; –0.083 (95 % CI [–0.131, –0.036]; $p = 6.3 \times 10^{-4}$) in the West; and –0.041 (95 % CI [–0.105, 0.022]; $p = 0.20$) in the East. The positive northern age slope likely reflects indoor exposure from older housing stock, whereas the southern decline aligns with accelerated metabolic clearance and reduced floor-contact behavior as children age. The HBM4EU children data suggest that childhood mbzp trajectories vary by region, reflecting differences in exposure sources and metabolic maturation, so modelling region-specific slopes could possibly capture this heterogeneity. Whether these divergent patterns persist into adolescence or for other biomarkers remains to be investigated.
Introducing the *sex*×age_c term increased AIC (ΔAIC = +2) and yielded a non-significant LR test ($\chi^2_1 = 0.01$, $p = 0.92$). Estimated age slopes in log(mbzp) were -0.009 yr^{-1} (95 % CI [−0.049, 0.031]; $p = 0.66$) for boys and -0.012 yr^{-1} (95 % CI [−0.052, 0.029]; $p = 0.58$) for girls, indicating no meaningful sex difference within this age band. This could be consistent with prepubertal children exhibiting comparable xenobiotic kinetics and exposure patterns, whereas hormonal changes in adolescence may introduce sex-specific trajectories.[15]

In the interaction model which includes both *region*×*season* and *region*×*DEGURBA*, two context-specific contrasts that remained statistically significant ($p < 0.05$) were extracted for interpretation: the winter–spring difference within each region with DEGURBA fixed at its reference level (urban) and the rural–urban difference within each region with season fixed at its reference level (spring). Estimates are presented on the log-mbzp scale with 95 % confidence intervals (Table 5).

Table 5: Statistically significant contrasts ($p < 0.05$) from Model 4 with interactions, including both *region*×*season* and *region*×*DEGURBA*. Estimates are on the log-mbzp scale; 95 % CIs.

Contrast	β	95 % CI	p
Winter – Spring (by region)			
East	–0.474	–0.743 – –0.205	0.001
Rural – Urban (by region)			
North	1.319	1.030 – 1.608	$< 2 \times 10^{-16}$
South	–0.392	–0.615 – –0.170	0.001
East	0.341	0.116 – 0.565	0.003

Each candidate model, whether or not it contained interaction terms, retained the full set of main-effect covariates (region, sampling season, DEGURBA, sex, centred age, and household education); interaction blocks were added on top of this common core. The fitted coefficients were projected onto the 288-cell EU-27 reference grid; delta-method SEs were calculated on the log-scale and back-transformed to obtain geometric means. Model 1, the main-effects specification (see also section

4.2.1), had an AIC of 8507 and an EU-standardised log-mean of $\widehat{\overline{Y}}^{(1)} = 1.4594$ (SE = 0.0063). Adding the *region*×*season* block (Model 2) lowered AIC by 27 but left the log-mean unchanged ($\widehat{\overline{Y}}^{(2)} = 1.4594$, SE = 0.0080). Adding the *region*×*DEGURBA* block alone (Model 3) gave a larger improvement (ΔAIC = –124) and raised the log-mean to 1.5140 (SE = 0.0075). Including both interaction blocks simultaneously (Model 4) produced the best fit (AIC = 8359; ΔAIC = –147.4; LR $\chi^2_{15} = 177.4$, $p < 2 \times 10^{-16}$) and an EU-standardised log-mean of 1.5000 (SE = 0.0090). Exponentiation yields the population-weighted geometric means and 95 % CIs reported in Table 6.

Table 6: EU-standardised mbzp means, log and concentration scales, from candidate OLS models projected onto the EU-27 reference grid.

Model	$\widehat{\overline{Y}}^{(m)}$ (log) ± SE_Δ	GM (µg g^{-1} crt)	95 % CI
1 (no interactions)	1.4594 ± 0.0063	4.30	4.25–4.36
2 (+ region×season)	1.4594 ± 0.0080	4.32	4.25–4.39
3 (+ region×DEGURBA)	1.5140 ± 0.0075	4.50	4.44–4.57
4 (+ region×season + region×DEGURBA)	1.5000 ± 0.0090	4.48	4.41–4.56

The different interaction specifications altered the EU-standardised log-mean by no more than ≈ 0.055 log-units (ca. 4–5 % on the $\mu g\, g^{-1}$ creatinine scale) relative to Model 1. Of the five two-way interactions tested, *region*×*season* and *region*×*DEGURBA* were retained as they captured geographic heterogeneity (large LR statistics and ΔAIC of –26 and –124, respectively). The terms *region*×*age*, and *sex*×*age* may prove relevant, especially in adolescent or adult populations, where age trajectories and pubertal physiology could become more prominent modifiers. However, they would add little information within the narrow 6–12 year EU-27 band (for which the age was fixed at 9 years, the midpoint of a nearly uniform distribution), and are therefore proposed for future work in other age groups and where age- and sex-specific trajectories may be more pronounced.

4.3 Mixed models-based standardisation

4.3.1 Diagnostics

For both **random-intercept fits, with and without interaction terms**, standardised conditional residuals (i.e., residuals after removing the estimated cohort effects) show no systematic curvature and no funnel-shaped pattern in the residual-versus-fitted plots (panels a, c of Fig. 16; Appendix); hence the homoscedasticity assumption appears credible. The accompanying QQ-plots (panels b, d of Fig. 16; Appendix) remain essentially linear over the central 75 % of the observations, with a few points in the extreme tails departing from the $N(0, 1)$ reference, indicating mild rather than consequential heavy-tailedness. Residual spread at high fitted values is slightly narrower in the interaction model, suggesting that the *region*×*season* and *region*×*DEGURBA* terms may absorb some residual heterogeneity. **For comparison, diagnostics were likewise produced for the ordinary–least–squares projections (Models 1 and 4 of Section 4.2; Fig. 17; Appendix)**. Their residual–versus–fitted plots only show a slight increase in spread at very low fitted values. The accompanying Q-Q plots again track the $N(0, 1)$ reference through roughly the central 75 % of the data, before displaying the same mild heavy-tailed pattern observed for the mixed models. Hence, the key **distributional assumptions of linearity, homoscedastic errors, and approximate normality appear reasonable for both the mixed-effects and the OLS specifications**.

To also verify that the two interaction blocks did not introduce problematic multicollinearity in the *fixed–effects* design matrix, the random intercept was temporarily ignored, fitting an OLS model with the same fixed terms, and computed **Generalised Variance Inflation Factors (GVIFs)**. GVIFs assess collinearity in the fixed-effects matrix X; the random-intercept term adds a separate Z-matrix and leaves X unchanged. Therefore, an OLS fit provides the correct X for collinearity diagnostics. The scaled indices were $\text{GVIF}^{1/(2\,\text{df})}$: Region 2.67 (df = 3); sampling-season 2.16 (df = 3); DEGURBA 2.62 (df = 2); sex 1.01 (df = 1); age_c 1.31 (df = 1); household-education 1.05 (df = 2); region×season 1.66 (df = 9); and region×DEGURBA 1.70 (df = 6). All values are well below the conventional thresholds of 4–5, indicating that multicollinearity among the fixed predictors is minor and unlikely to compromise the interaction-model estimates.

The **cohort-intercept BLUPs** (Fig. 18; Appendix) were centered at zero and closely follow a $\mathcal{N}(0, \hat{\sigma}_u^2)$ shape ($\hat{\sigma}_u = 0.485$ in the baseline, 0.472 in the with interactions mixed models), with almost all values between -1 and $+1$ on the log scale and only a hint of heavier tails. Therefore, the Gaussian random-effects assumption appears reasonable, and no single cohort dominates the fixed-effect estimates. **Cohort-level Cook's distances** mostly fall below the conventional cut-off $D > 4/n \approx 0.36$ (see Figure 19; Appendix). Only cohort 4 (NEBII–NO; $D \approx 0.46$) and cohort 10 (OCC–DK; $D \approx 0.43$) exceeded the threshold. A leave-one-out **sensitivity analysis** of the baseline mixed model produced $\mu_{\log} = 1.39$ (0.0387) with all cohorts; 1.32 (0.0321) after removing cohort 4; 1.48 (0.0315) after removing cohort 10; and 1.45 (0.0363) when both 4 & 10 were omitted (SEs in parentheses; Δ-method). These shifts of about ±0.07–0.09 log-units are larger than the Δ-method SE yet still well below the Monte-Carlo SDs reported in Table 8. Hence, while cohorts 4 and 10 are the most influential, their impact might be modest relative to the overall Monte-Carlo uncertainty. All 11 cohorts were retained, with cohorts 4 and 10 flagged as influential.

4.3.2 Mixed models fitting; EU-27 standardised values

For the random-intercept mixed model without interactions the between–cohort variance was: $\hat{\sigma}_u^2 = 0.235$ ($\hat{\sigma}_u = 0.485$) and the within-cohort (residual) variance $\hat{\sigma}_e^2 = 1.103$ ($\hat{\sigma}_e = 1.050$), giving an intraclass–correlation:

$$\widehat{\mathrm{ICC}} = \frac{0.235}{0.235 + 1.103} = 0.176.$$

Thus 17.6 % of the variability in log-mbzp (standardised for creatinine) is attributable to differences between cohorts. In the mixed model with interactions the corresponding estimates, $\hat{\sigma}_u^2 = 0.223$ ($\hat{\sigma}_u = 0.472$) and $\hat{\sigma}_e^2 = 1.082$ ($\hat{\sigma}_e = 1.040$), yield $\widehat{\mathrm{ICC}} = 0.171$; hence 17.1 % of the variance is attributable between cohorts. Moreover, for the latter, fit improved with AIC dropping from 8063.7 to 8040 (mixed effects model without interactions) and a likelihood-ratio test ($\chi^2_{15} = 53.7$, $p = 2.9 \times 10^{-6}$).

Table 7: Fixed–effect estimates (log-mbzp scale) from random-intercept mixed models.

Effect	**Mixed baseline model**		**Mixed interaction model**	
	Estimate	SE	Estimate	SE
age_c	−0.1056	0.0180	−0.1120	0.0180
Additional interaction effects				
$\mathrm{region}_{\mathrm{South}}$:summer	—	—	0.4775	0.2098
$\mathrm{region}_{\mathrm{West}}$:summer	—	—	0.4263	0.1752
$\mathrm{region}_{\mathrm{East}}$:winter	—	—	−0.4730	0.1818
$\mathrm{degurba}_{\mathrm{Rural}}$	—	—	0.5122	0.1648
$\mathrm{region}_{\mathrm{South}}$:$\mathrm{degurba}_{\mathrm{Rural}}$	—	—	−0.9313	0.2220
$\mathrm{region}_{\mathrm{East}}$:$\mathrm{degurba}_{\mathrm{Rural}}$	—	—	−0.6469	0.2430

Note. Only effects with $p < 0.05$ are displayed. The baseline specification contains only main-effect terms; the interaction specification additionally includes *region×season* and *region×DEGURBA*. Reference levels: North region, Spring season, Urban DEGURBA, male sex, low household education. Cohort is modelled as a random intercept.

A negative association with age persisted in both random–intercept specifications on the log-mzp scale, implying an ≈ 10% reduction in mbzp concentration for each additional year [$\exp(-0.1056) \approx 0.90$. The interaction model, however, disclosed geographically specific modifications (see Table 7): summer concentrations were higher in the *South* ($\hat{\beta} = 0.48$) and *West* ($\hat{\beta} = 0.43$), whereas winter values were lower in the *East* ($\hat{\beta} = -0.47$). Urban–rural differences also varied by region; in the South the overall rural excess of +0.51 log-units was counteracted by a South×Rural interaction of −0.93, yielding a net estimate of about −0.42 log-units for rural children in that region. The *baseline* mixed model used the same fixed covariates as the OLS main–effects model (Table 3) but added a random intercept for cohort. Once this between-cohort variance is absorbed, the large regional main effects (South > North, West > North, East < North) that were highly significant in OLS lose significance and disappear from Table 7. Hence those broad regional gaps arise chiefly

between cohorts, not within them. Moreover, the mixed *with interactions* model retains the random intercept and introduces *region×season* and *region×DEGURBA* blocks. The context-specific terms—South:summer, West:summer, East:winter, South:Rural, etc.—match in sign and magnitude the significant contrasts identified by the OLS with interactions model (Table 5), agreeing to within ≈ 0.03 log-units and sharing the same 95 % confidence bounds. Therefore, after cohort-level heterogeneity is accounted for, only the fine-grained *region-specific seasonal* and *urban–rural* contrasts remain important.

Table 8: EU–standardised log-mbzp means, geometric means from random-intercept mixed models

Model	Method	$\hat{\mu}_{\log}$	SE/SD[a]	$\exp(\hat{\mu})$	95% CI (orig.)
Baseline	Delta	1.391	0.0387	4.02	[3.73, 4.34]
	MC–fixed	1.389	0.1620	4.01	[2.92, 5.51]
	MC–full	1.385	0.5133	3.99	[1.46, 10.9]
+ 2-way inter.	Delta	1.376	0.0383	3.96	[3.67, 4.27]
	MC–fixed	1.378	0.1582	3.96	[2.91, 5.41]
	MC–full	1.380	0.4980	3.97	[1.50, 10.6]

[a] SE for the Delta method; Monte-Carlo SD for the two simulation-based schemes ($M = 5{,}000$ replicate means). Concentrations at µg g^{-1} creatinine

Table 8 presents the EU-27–standardised means and three uncertainty estimates for the two random-intercept models. For the **baseline** mixed model, projecting onto the EU-27 reference grid gave: $\hat{\mu}_{\Delta} = 1.391$ (0.039), $\hat{\mu}_{\mathrm{MC,fix}} = 1.389$ (0.162) and $\hat{\mu}_{\mathrm{MC,full}} = 1.385$ (0.513). For the **interaction** model the corresponding values were 1.376 (0.038), 1.378 (0.158) and 1.380 (0.498). Back-transforming puts both models at $\approx 4\ \mu\mathrm{g\,g}^{-1}$ creatinine, but the 95 % limits widen from the Delta estimate to MC-fixed and again to MC-full, reflecting the successive inclusion of fixed-effect sampling error and, finally, between-cohort heterogeneity.

Switching from the Delta method ($\mathrm{SE}_{\Delta} = 0.039$) to the MC–fixed procedure ($\widehat{\mathrm{SD}}_{\mathrm{fix}} = 0.162$) increases the uncertainty by roughly a factor of four. In the MC–fixed scheme, each Monte-Carlo replicate applies a single draw of the coefficient vector $\boldsymbol{\beta}$ to every one of the $G = 288$ grid cells; the replicate means therefore inherit the entire sampling variance of $\boldsymbol{\beta}$ and treat all cells as perfectly correlated. By contrast, the Delta method first computes the variance of the prediction in *each* cell and only then averages them, down–weighting by the squared calibration weights $w_g^2 \approx (1/288)^2 \simeq 1.2 \times 10^{-5}$. Consequently, the contribution of any single cell to the overall variance is on the order of 0.001%, and the cell-specific uncertainties largely cancel:

$$\mathrm{Var}(\hat{\mu}_{\Delta}) = \sum_{g=1}^{G} w_g^2 \Big[\mathbf{x}_g^{\top} \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) \mathbf{x}_g + \hat{\sigma}_u^2 \Big].$$

Because the MC–fixed replicates do not benefit from this w_g^2-attenuation, their SD is appreciably larger, while the MC–full variant inflates the uncertainty further by adding a random draw from

the cohort-level variance component $\hat{\sigma}_u^2$ to every replicate. More specifically, when the cohort–level random intercept is resampled as well (MC–full), the uncertainty widens by a further factor of ≈ 3, giving $\widehat{\mathrm{SD}}_{\mathrm{full}} = 0.513$. This arises because the cohort variance $\hat{\sigma}_u^2$ now enters the variance of the EU-standardised mean *once*, rather than being multiplied by the weight-squared term $\sum_g w_g^2$ that dilutes it in the Delta expression. Since both Monte-Carlo schemes aggregate the simulated log-means *before* applying the external weights, their standard deviations must exceed the Delta SE. Introducing the *region*×*season* and *region*×*DEGURBA* blocks reduces $\widehat{\mathrm{SD}}_{\mathrm{full}}$ only marginally (0.513 $\rightarrow$ 0.498), leaving the MC–full 95 % interval roughly an order of magnitude wider than the Delta interval. MC–full therefore provides the most conservative confidence limits, with the interactions trimming those limits only slightly by absorbing context-specific contrasts that would otherwise contribute to unexplained between-cohort heterogeneity.

Moreover, the SEs rised sharply when we moved from the OLS grid projection (Section 4.2) to the random-intercept mixed-effects projection hereby. Some of the potential reasons are outlined below:

(i) Explicit between–cohort variance: The mixed model treats cohort as a random factor, so the estimated variance component $\hat{\sigma}_u^2 > 0$ is propagated to every grid cell; OLS implicitly fixes $\sigma_u^2 = 0$.

ii) Uncertainty in the empirical-Bayes intercepts: Each cohort effect is estimated by a BLUP,

$$\hat{b}_j \;=\; \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_j}\,(\bar{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}),$$

where n_j is the number of children in cohort j. The shrinkage factor pulls the cohort mean $\bar{Y}_j$ toward the fitted grand mean $\mathbf{X}_j\hat{\boldsymbol{\beta}}$ more aggressively when the cohort is small (n_j low) or noisy ($\hat{\sigma}_e^2$ large). Although this stabilises the point estimate, $\hat{b}_j$ remains a random quantity, and its sampling variance must still be propagated.

iii) Correlated resampling in MC–full: In the MC–full scheme a single draw $u^{(m)} \sim \mathcal{N}(0, \hat{\sigma}_u^2)$ is added to the linear predictor of *all* $G = 288$ grid cells within replicate m. Because the same random intercept is applied everywhere, the replicate means are perfectly correlated across cells, inflating the replicate-to-replicate spread.

The combined impact of these three mechanisms—an explicit between-cohort variance, uncertainty in the shrinkage estimates, and joint resampling of the random intercept—raises the SEs by an order of magnitude relative to the OLS projection, while potentially providing a more realistic reflection of uncertainty for data that are clustered across multiple cohorts and countries.

4.4 Non-model, design-based methods

4.4.1 Direct post-stratification weighting

An **unweighted mean** of the log-transformed, creatinine-standardised mbzp concentration was first calculated on the biomarker-complete subset (n = 2 784), yielding a geometric mean of:

$$\exp(1.2616) \;=\; 3.53\ \mu g\, g^{-1}\, crt$$

Restricting further to the n = 2 722 children, with complete stratification data as well, it changed the log-mean only slightly to 1.2638, corresponding to:

$$\exp(1.2638) = 3.54\ \mu g\, g^{-1}\, \text{crt},$$

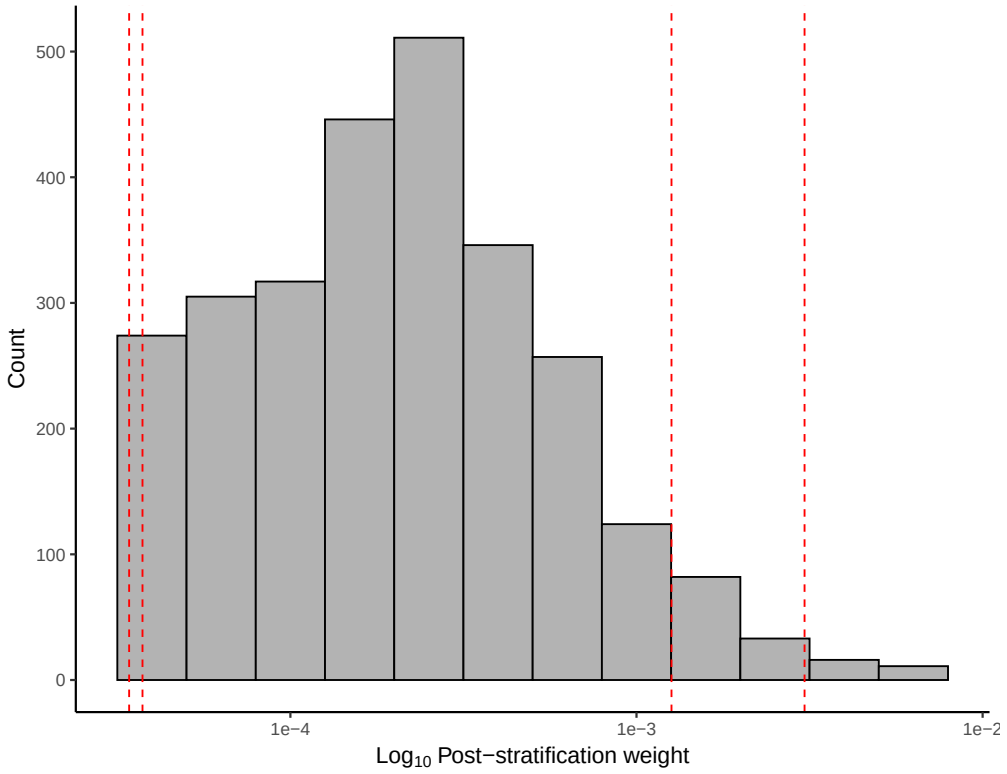


Figure 13: Histogram of $\log_{10}$ post-stratification *probability* weights. Grey bars show the count per 0.2-dex bin; red dashed lines mark the 1st, 5th, 95th and 99th percentiles.

After this restriction, direct post-stratification weights were computed by matching each of the 2 722 observations to the Eurostat reference grid (see Section 3.2), dividing each cell's overall weight by its sample count, and renormalising to sum to one. These observation-level weights ranged from 3.42×10^{-5} to 6.63×10^{-3} (median 2.00×10^{-4}; IQR 8.88×10^{-5}–3.95×10^{-4}). This distribution indicates modest heterogeneity in the post-stratification adjustment; most weights lie within a 2- to 5-fold range around the median and show no extreme outliers that would unduly influence the weighted mean (Figure 13).

The **direct postratification weighted mean** of log-mbzp corresponded to a geometric mean of:

$$\exp(1.5671) \;=\; 4.79\ \mu\mathrm{g\,g^{-1}}\,crt$$

Thus, post-stratification increased the geometric mean by 35.6 % relative to the unweighted estimate. Imputing the 62 missing ISCED values to "Medium" and recomputing the weights left the log-mean unchanged at 1.567, indicating negligible impact of the missing stratification data.

4.4.2 Survey-design weighted estimation; with and without clustering

Direct post-stratification weighting provides unbiased point estimates of the population mean (see section 4.4.1), but no analytic variance. The **survey-design framework** overcomes this by treating the weights as inverse inclusion probabilities and applying Taylor-series linearization to derive design-based standard errors (SEs) and confidence intervals (CIs).

Independent design - no clustering: Assuming an *independent* design—i.e. every child treated as its own primary sampling unit, the log-scale SE was estimated at 0.048 for a log-scale mean = 1.5671 and a corresponding EU-standardised geometric mean of:

$$4.79\ \mu\mathrm{g\,g^{-1}}\,\mathrm{crt}\ [95\ \%\ \mathrm{CI}\text{: }4.37\text{–}5.26\ \mu\mathrm{g\ g^{-1}\,crt}]$$

Weights span two orders of magnitude (3.4×10^{-5}–6.6×10^{-3}; max / min $\approx$ 194); max/min ca. 194), yet the weight-only design effect appears modest (DEFF = 3.56), leaving an effective sample of $n_{\mathrm{eff}} \approx 764$. This may suggest that variance inflation is driven by the systematic post-stratification adjustment across many cells rather than by a few outlying weights. Region-specific DEFFs were highest in the West (0.55), followed by the South (0.23) and East (0.11), and lowest in the North (0.02), indicating that variance inflation is greatest in strata where the study sample departs most strongly from the EU-27 population distribution.

Weight-trimming sensitivity: Capping the heaviest 1 % of weights (99^{th} percentile, 0.00306) reduced the geometric mean to 4.67 µg $\mathrm{g^{-1}}$ crt (95 % CI 4.33–5.03), a –2.6 % change, inside the original CI. A 5 % cap (0.00126) lowered the estimate to 4.51 µg $\mathrm{g^{-1}}$ crt (95 % CI 4.22–4.81), a –6.0 % shift that likewise remains within the untrimmed CI. Hence extreme weights appear to exert only modest influence, and the untrimmed estimate could be retained.

Clustering: Specifying cohort as the sampling cluster left the point estimate unchanged (1.5671) but raised the log-scale SE to 0.28, widening the interval to:

$$4.79\ \mu\mathrm{g\,g^{-1}}\,\mathrm{crt}\ [95\ \%\ \mathrm{CI}\text{: }2.76\text{–}8.32\ \mu\mathrm{g\ g^{-1}\,crt}]$$

Clustering by cohort increases the log-scale SE from 0.048 to 0.280 (a factor of 5.8). Thereby, intra-cohort correlation, not extreme weights, is the principal source of uncertainty: the independent-design CI slightly understates true variability, whereas the cluster-robust interval provides a more conservative precision estimate for multi-site data.

Sensitivity calibration for single-year age. Given that Eurostat data suggested an almost even split of 6–12-year-olds across single ages, within every EU-27 region (Section 3.1), the existing five-way post-stratification *frequency* weights were recalibrated to a synthetic *region* × *age* margin that assigned one-seventh of each region's weight to each age from 6 to 12 years. Recalibration was first performed with `rake()` on the clustered design (cohorts as PSUs). As the HBM4EU sample itself is not uniformly distributed across single years within regions (see Section 4.1), the additional constraint required substantial weight adjustments. The EU-standardised geometric mean dropped to:

$$3.33\ \mu g\,g^{-1}\,\mathrm{crt} \quad [95\%\ \mathrm{CI}: 2.54\text{–}4.36],$$

a 30 % decrease relative to the main clustered estimate of 4.79 µg g^{-1} crt. It should be highlighted that the weights-only design effect rose from 3.14 to 5.29, thus reducing the effective sample size substantially from $n_{\mathrm{eff}} \approx 867$ to $n_{\mathrm{eff}} \approx 515$.

Repeating the calibration without cohort clustering, it yielded an almost identical geometric mean and a narrower CI:

$$3.33\ \mu g\,g^{-1}\,\mathrm{crt} \quad [95\%\ \mathrm{CI}: 3.06\text{–}3.62],$$

The additional age margin changed the point estimate substantially (–30 %); albeit it enlarged the weights-only design effect from 3.14 to 5.29 (n_{eff}: 867 → 515), with the precision dropping sharply. The six-way calibration was kept only as a sensitivity check, and the five-way weights remain the primary design for all analysis and the following section as well.

4.4.3 Raking (marginal calibration)

Raking was applied to the $n = 2\,722$ children with complete biomarker and auxiliary data, aligning the sample with the *marginal* Eurostat distributions of region, sex, sampling season, DEGURBA and household ISCED while preserving the weight total $\sum_i \omega_i = n$. On the probability scale the calibrated weights ranged from 5.1×10^{-5} to 6.5×10^{-3} (median 2.5×10^{-4}; IQR 1.2×10^{-4}–4.0×10^{-4}) (Figure 14). These correspond to 0.14–17.72 on the frequency scale ($\Sigma\omega_i = n$) and form a tighter distribution than the full five-way post-stratification weights (Figure 13). The weights-only design effect was DEFF = 3.14, giving an effective sample size $n_{\text{eff}} \approx 867$. Relative to the post-stratified weights (see Section 4.4.2), raking lowered DEFF by about 12 %, indicating a modest gain in precision without trimming or capping any weights.

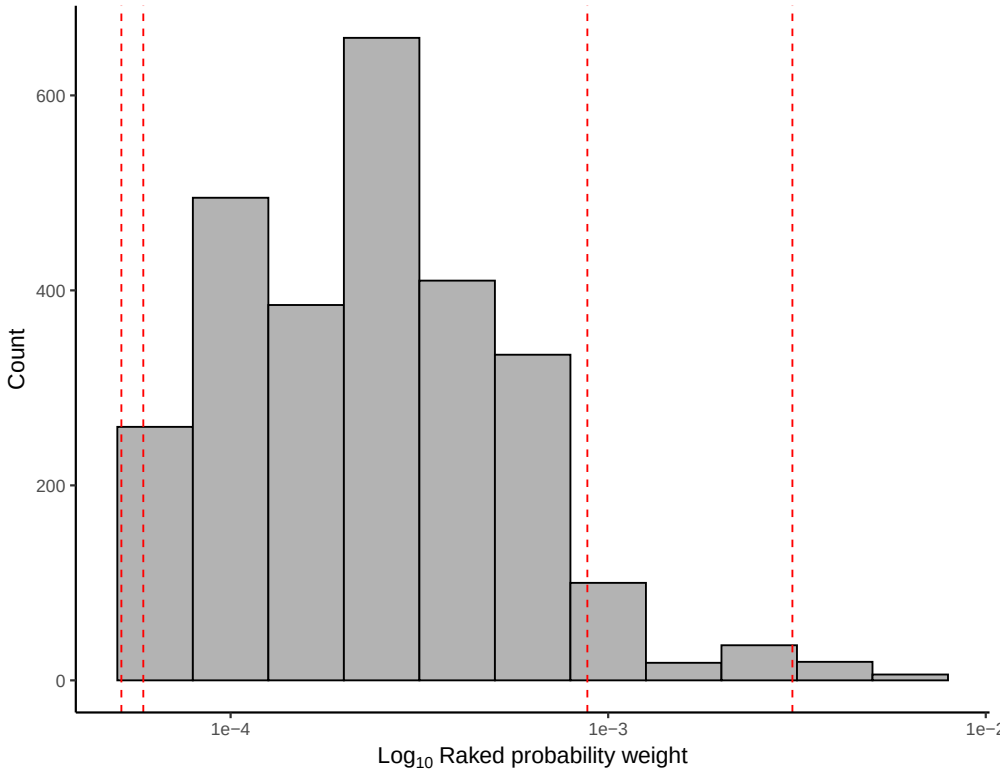


Figure 14: Histogram of $\log_{10}$ raked *probability* weights. Grey bars show the count per 0.2-dex bin; red dashed lines mark the 1st, 5th, 95th and 99th percentiles.

The raked log-scale estimate and its SE were estimated as $\hat{\mu}_{\text{rake}} = 1.5247$ (SE = 0.0421). Back-transformation yielded the EU-standardised geometric mean:

$$4.59\ \mu\text{g}\,\text{g}^{-1}\,\text{crt} \quad [95\%\ \text{CI: } 4.23\text{–}4.99].$$

This value is 4.2 % lower than the post-stratified mean (4.79 µg g^{-1} crt), yet lies inside its 95 % CI (4.37–5.26 µg g^{-1} crt), showing that marginal calibration leaves the central estimate essentially unchanged.

Clustering by cohort: Treating cohort as the PSU (`ids = ~cohort`) left the point estimate the same but widened the 95 % CI to 2.77–7.62 µg g^{-1} crt (log-scale SE = 0.2580), suggesting, also here, that intra-cohort correlation, drives the sampling variance.

Overall, raking appears to smooth the weight distribution and trims the design effect while preserving the EU-standardised geometric mean. It therefore offers a modest efficiency gain and could be an alternative when fully post-stratified weights are highly variable or unstable.

5 Overview, Conclusions and Recommendations

5.1 Overview of estimation results - conclusions

Table 9 shows a tighter spread of EU-standardised/ post-stratified geometric means across all non-clustered estimators. The direct post-stratified mean defines the upper bound, whereas the random-intercept mixed model with analytic Δ SE defines the lower bound. Declaring cohorts as primary sampling units widens substantially CIs, implying that between-study correlation may dominate sampling variance. For the mixed-effects models, the three uncertainty estimation schemes behave as expected: the analytic Δ method is computationally light, yielding the narrowest interval; MC-fixed adds fixed-effect sampling error and widens the interval moderately; MC-full further injects a random cohort effect in every replicate, producing the broadest (and most conservative) limits. The choice of SE method influences precision rather than central tendency. Weight trimming or marginal raking smooth the extreme tails of the weight distribution and shift the mean only marginally. In contrast, adding a region $\times$ age calibration margin lowered the mean and almost halved the effective sample size, illustrating the bias–variance trade-off when additional population constraints are imposed. Introducing biologically plausible two-way interactions (e.g. hereby region $\times$ season, region $\times$ DEGURBA) in either OLS or mixed-effects frameworks leaves both means and intervals very similar, indicating that any residual effect modification is minor relative to the main-effect structure already modelled.

Overall, the mixed-effects model-based direct standardisation (with MC-fixed interval) may offer a midpoint that recognises cohort heterogeneity; while the design-based clustered intervals provide a conservative outer envelope. When a design-based alternative is preferred—or model fit proves unstable—marginal raking could be a viable substitute: it smooths the heaviest weights and reduces design effects relative to full post-stratification. Discussion on suggestions for future research and for alternative methodologies is provided under section 5.3.

Table 9: EU-27 geometric mean (µg g^{-1} crt) for mbzp_impcrtlog across all estimation strategies

Type	Estimator	SE/SD method	GM	95 % CI	Comment
Descriptive (no weighting)					
	Unweighted mean	–	3.53	–	HBM4EU sample-based
Model-based direct standardisation (OLS; EU-27 grid)					
	OLS, Model 1	Δ	4.30	4.25–4.36	no interaction terms
	OLS, Model 4	Δ	4.48	4.41–4.56	+ 2-way interactions
Model-based direct standardisation (random-intercept mixed effects; EU-27 grid)					
		Δ	4.02	3.73–4.34	no interaction terms
	No interaction terms	MC-fixed	4.01	2.92–5.51	no interaction terms
		MC-full	3.99	1.46–10.9	no interaction terms
		Δ	3.96	3.67–4.27	+ 2-way interactions
	with interactions	MC-fixed	3.96	2.91–5.41	+ 2-way interactions
		MC-full	3.97	1.50–10.6	+ 2-way interactions
Design-based (5-way strata)					
	Post-strat., independent	–	4.79	4.37–5.26	DEFF 3.56; $n_{\text{eff}} \approx 764$; 1% trim → 4.67, 5% trim → 4.51
	Post-strat., cluster PSU	–	4.79	2.76–8.32	cohorts as PSUs
	Raked, independent	–	4.59	4.23–4.99	DEFF 3.14; $n_{\text{eff}} \approx 867$
	Raked, cluster PSU	–	4.59	2.77–7.62	cohorts as PSUs
Design-based (6-way strata, with region × age calibration)					
	Calibrated, independent	–	3.33	3.06–3.62	-
	Calibrated, cluster PSU	–	3.33	2.54–4.36	cohorts as PSUs; DEFF 5.29; $n_{\text{eff}} \approx 515$

[a] interactions refer to both 2-way interactions: *region×season* and *region×DEGURBA*.
[b] Five-way stratification uses *region, sex, sampling season, DEGURBA* and *household ISCED*. Δ = analytic Delta-method SE; MC-fixed = Monte-Carlo SE that propagates fixed-effect uncertainty only; MC-full = Monte-Carlo SE that additionally draws one cohort-level random intercept u in each replicate. GM = geometric mean on the original concentration scale.

5.2 Methodological comparison: opportunities and limitations

Model-based approaches (A–B) hold on the transportability assumption: the covariate effects estimated for the biomarker concentration in the HBM4EU sample remain valid for every demographic profile represented in the EU-27 reference grid. **Design-based methods** (C–E) assume only that, once weighted, the sample represents the EU-27 population within each post-stratum; no further model-based extrapolation is required.

A. Model-based direct standardisation (OLS; EU-27 grid)

- Fits OLS to log(mbzp/crt); optional 2-way interactions.
- Predicts all reference grid cells and aggregates with EU-27 weights.
- **Limits:** linearity, homoscedastic errors, no between-cohort variance (σ_u^2=0); Delta SE ignores weight and imputation uncertainty.

B. Model-based direct standardisation (random-intercept mixed model; EU-27 grid)

- Adds a cohort-level random intercept (u); optional two-way interactions.
- SE estimation: Δ (analytic); MC-fixed (propagates fixed-effect error); MC-full (propagates fixed-effect *plus* random-intercept error). Uncertainty in the calibration weights still ignored.
- *Limits:* assumes normally distributed u and no random slopes; all SE methods treat $\hat{\sigma}_u^2$ as fixed; MC-full yields the widest interval and is computationally slower because the same random intercept is applied to every grid cell.

C. Direct post-stratification (design-based)

- Re-weights every child to its five-way stratum; makes no parametric assumptions.
- *Limits:* variance must be obtained by Taylor linearisation or replicate weights; cohort clustering is ignored; sparse cells inflate weights; cannot predict for strata absent from the sample.

D. Survey-design estimation (post-stratified weights)

- Uses the same five-way post-stratification weights as in (C), but embeds them in a `survey` design object. The analyst can specify either an *independent* design or a *clustered* design (cohort as the primary sampling unit) and must report the design effect (DEFF) and the implied effective sample size n_{eff}.
- *Limits:* treating cohorts as PSUs can inflate confidence intervals several-fold; adding further margins or trimming extreme weights alters the DEFF and may erode precision; no parametric model is fitted, so interaction tests are unavailable; SEs still omit uncertainty in the calibration margins and in any imputed values.

E. Raking calibration (marginal weighting)

- Iteratively adjusts the original post-stratification weights so the weighted sample matches each Eurostat *marginal* (region, sex, season, DEGURBA, ISCED). This typically lowers the design effect and tames extreme weights without ad-hoc trimming; the new weights can be analysed directly, with or without cohort-clustered SEs.
- *Limits:* aligns single margins only—unmodelled interactions may leave residual bias; convergence can fail or inflate weights if some cells are very sparse; SEs still ignore uncertainty in the Eurostat margins and any imputation.

5.3 Ideas for future research - EU level reference values

To further investigate obtaining EU-level reference values that are both design-consistent and policy-relevant in forthcoming biomonitoring initiatives (e.g., the PARC-aligned studies), the following methodological recommendations or further research suggestions are given:

- **Provision of cohort-specific survey design weights *prior* to data harmonisation.** Each contributing study should supply probability (or post-stratification) weights before pooling, so that all downstream estimates remain design-consistent.
- **Dual reporting of model-based and design-based estimates and intervals.** For example, mixed-effects direct standardisation (with MC-fixed SEs) can be presented as an efficient, cohort-aware estimate, while a raked, cluster-robust survey design could offer an assumption check and a conservative envelope.
- **Adopt finer margins, where feasible.** Depending on data availability, Eurostat *country*-level proportions (rather than broad regional totals) could be explored in the weighting grid to correct for countries that are over- or under-represented in HBM4EU; this may improve the accuracy of the EU-level reference value without aiming at country-specific estimates.
- **Use joint margins if available.** If Eurostat may provide cross-tabulations, iterative proportional fitting could be applied, removing residual interaction bias without ad-hoc trimming.
- **Age calibration and broader age coverage.** Age–specific counts (per region) for both sexes were even within the 6–12 yr span and an centered-age reference grid was used hereby. When regionxage was added as an additional calibration margin under the survey-design methods, it brought negligible loss of precision. One should further explore the inclusion of age margins also to the other age groups: teenagers and adults, where hormonal changes and behaviour may create sex-specific, age-dependent exposure trajectories that the current children-only grid cannot capture.
- **Alternative modelling strategies.** Two complementary routes can be pursued:

  (i) *Design-based GLMs* (`svyglm`). Providing EU-level marginal coefficients with sandwich standard errors that reflect both calibration weights and cohort clustering—no cross-population "transportability" assumption is required. *Limits:* cannot separate between- and within-cohort variance; precision drops if a few extreme weights dominate.

  (ii) *Generalised estimating equations (GEE / GEE2).* Yield comparable population-average effects while relaxing the independence-of-PSU assumption; sandwich SEs stay valid even when the working correlation is misspecified. *Limits:* cluster-level variance components and BLUPs are unavailable, very small PSUs inflate SEs, and efficiency falls when the chosen correlation departs notably from reality.

References

[1] World Health Organization Regional Office for Europe. *Human biomonitoring for Europe*. reference number: WHO/EURO:2023-7574-47341-69480

[2] Zare Jeddi, M., et al. *Developing Human Biomonitoring as a 21st Century Toolbox within the European Exposure Science Strategy 2020-2030*. Environment International, 168, 2022, p. 107476. DOI: https://doi.org/10.1016/j.envint.2022.107476. JRC129182.

[3] European Parliament and Council of the European Union. *Proposal for a Regulation establishing a common data platform on chemicals, laying down rules to ensure that the data contained in it are FAIR and establishing a monitoring and outlook framework for chemicals*. COM(2023) 779 final, 2023/0453 (COD), Brussels, 7 December 2023. SWD(2023) 855 final.

[4] Gilles, Liese, et al. *Harmonization of human biomonitoring studies in Europe: characteristics of the HBM4EU-aligned studies participants. International Journal of Environmental Research and Public Health*, 19(11), 6787, 2022. DOI: https://doi.org/10.3390/ijerph19116787.

[5] Ganzleben, C., et al. *Human biomonitoring as a tool to support chemicals regulation in the European Union. International Journal of Hygiene and Environmental Health*, 220, 94–97, 2017. DOI: https://doi.org/10.1016/j.ijheh.2016.09.007.

[6] Govarts, Eva, et al. *Harmonized human biomonitoring in European children, teenagers and adults: EU-wide exposure data of 11 chemical substance groups from the HBM4EU Aligned Studies (2014–2021). International Journal of Hygiene and Environmental Health*, 249, 2023, p. 114119. DOI: https://doi.org/10.1016/j.ijheh.2023.114119.

[7] Ougier, Eva, et al. *Chemical prioritisation strategy in the European human biomonitoring initiative (HBM4EU) – development and results. International Journal of Hygiene and Environmental Health*, 236, 2021, p. 113778. DOI: https://doi.org/10.1016/j.ijheh.2021.113778.

[8] Vorkamp, K., et al. *Biomarkers, matrices and analytical methods targeting human exposure to chemicals selected for a European human biomonitoring initiative. Environment International*, 146, 2021, 106082. DOI: https://doi.org/10.1016/j.envint.2020.106082.

[9] European Commission. *IPCHEM - Information Platform for Chemical Monitoring*, HBM4EU metadata Available online: https://ipchem.jrc.ec.europa.eu/

[10] Eurostat. *Population on 1 January by age and sex, data extracted for: 2023*. Available at: https://doi.org/10.2908/demo_pjan.

[11] Eurostat. *Distribution of population by degree of urbanisation, dwelling type and income group — Custom extract for 2020*. Available at: https://ec.europa.eu/eurostat/databrowser/view/ILC_LVHO01__custom_16774548/default/table?lang=en.

[12] Eurostat. *Distribution of households by educational attainment level of the reference person - experimental statistics — Custom extract for 2020*. Available at: https://ec.europa.eu/eurostat/databrowser/view/icw_car_03/default/table?lang=en.

[13] O'Brien, K. M., Upson, K., Cook, N. R., and Weinberg, C. R. *Environmental chemicals in urine and blood: improving methods for creatinine and lipid adjustment. Environmental Health Perspectives*, 124(2), 2016, pp. 220–227. DOI: https://doi.org/10.1289/ehp.1509693.

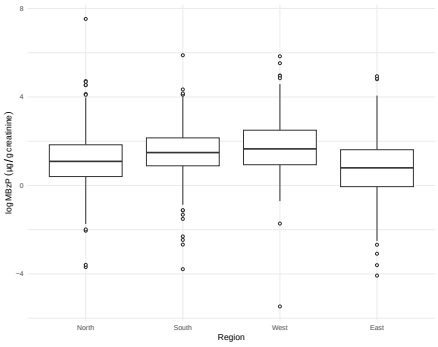
[14] Barr, D. B., Wilder, L. C., Caudill, S. P., Gonzalez, A. J., Needham, L. L., & Pirkle, J. L. Urinary Creatinine Concentrations in the U.S. Population: Implications for Urinary Biologic Monitoring Measurements. *Environmental Health Perspectives*, 113(2), 192–200 (2005). DOI: https://doi.org/10.1289/ehp.7337.

[15] Hines, R. N. Ontogeny of Human Hepatic Cytochrome P450 Enzymes: Impact on Drug Metabolism and Pharmacokinetics. *Pharmacology & Therapeutics*, 118(2), 250–267 (2008). DOI: https://doi.org/10.1016/j.pharmthera.2008.02.005.

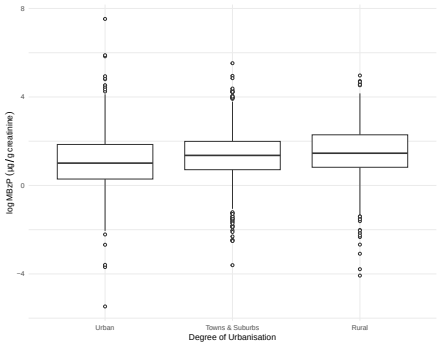
[16] Lumley T. Analysis of Complex Survey Samples. *Journal of Statistical Software* **9**(1), 1–19 (2004). Available at: https://doi.org/10.18637/jss.v009.i08.

Appendix

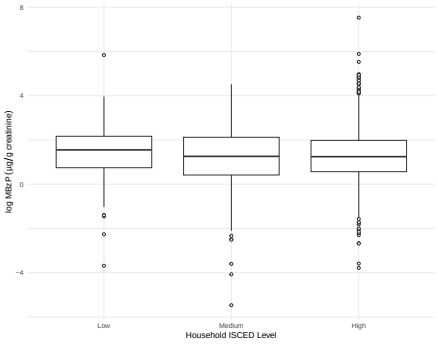
A Additional graphs



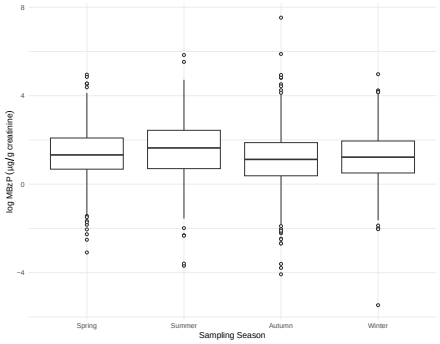
(a) By region



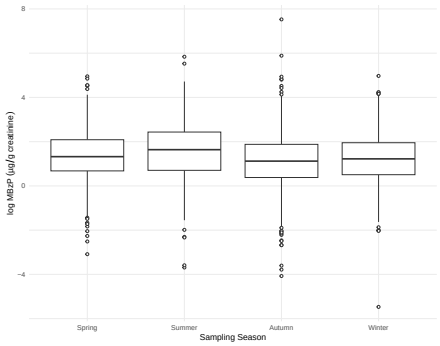
(b) By DEGURBA



(c) By ISCED



(d) By season



(e) By sex

Figure 15: Boxplots of log-mbzp (µg/g creatinine) across key strata - HBM4EU: children

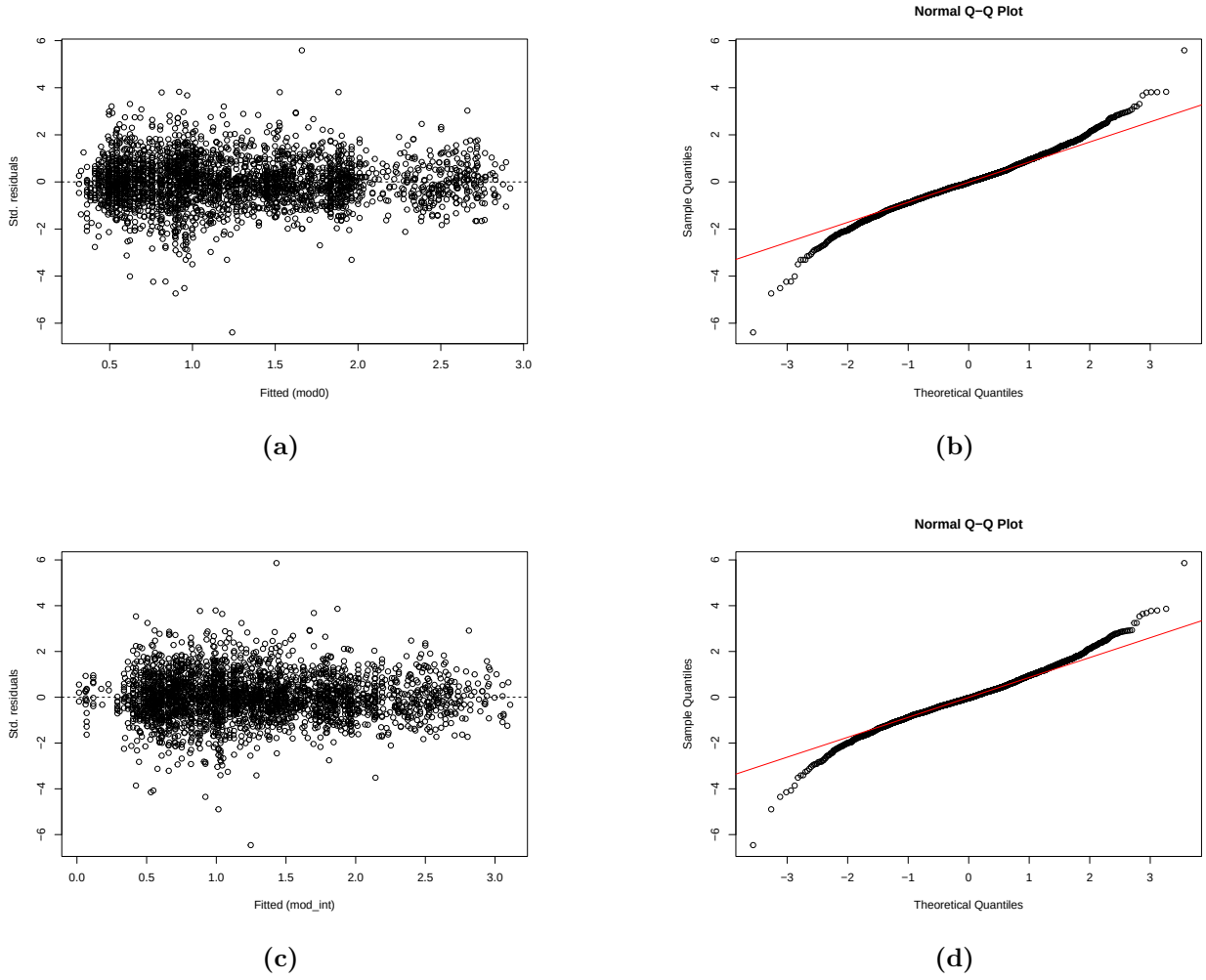


Figure 16: Diagnostic plots for the baseline (panels a–b) and with 2-way interactions (panels c–d) random-intercept models (see section 4.3). Panel (a) and (c) show standardized conditional residuals versus fitted values; panels (b) and (d) show QQ-plots against a $N(0, 1)$ reference.

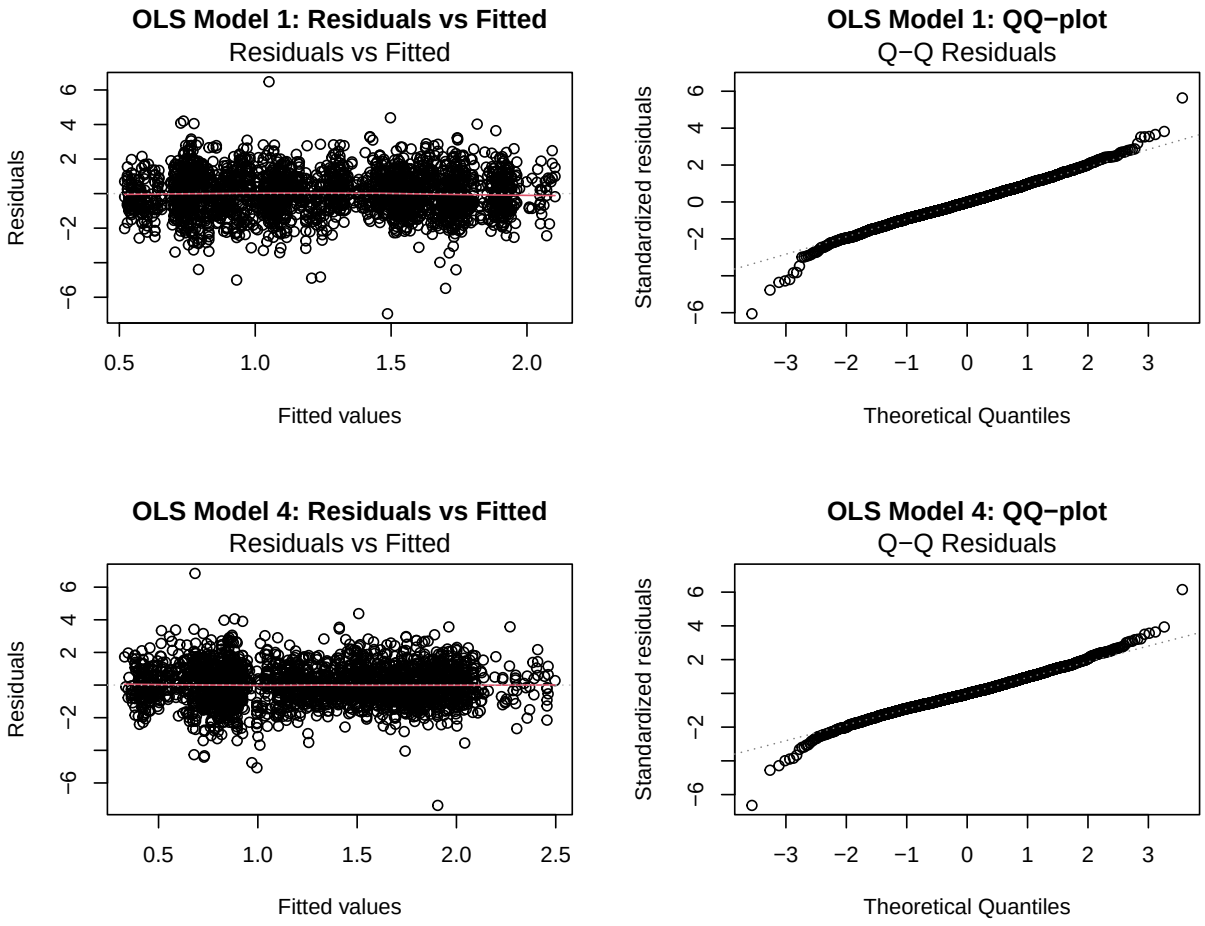
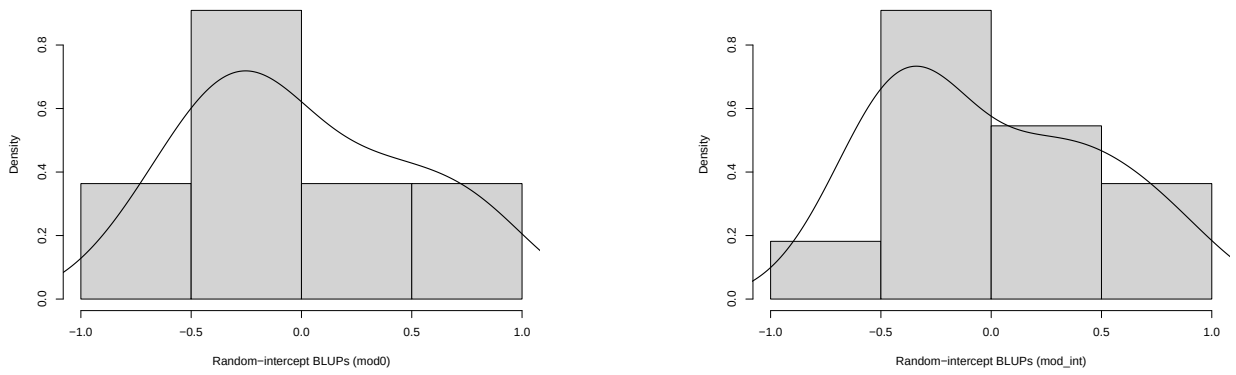


Figure 17: Diagnostic plots for the ordinary-least-squares models (see section 4.2). Top row: Model 1 (main effects only). Bottom row: Model 4 (region × season and region × DEGURBA interactions). Each row shows standardized residuals vs. fitted values (left) and the corresponding QQ-plot (right).



(a) **(b)**

Figure 18: Distribution of cohort–level BLUPs for the baseline (a) and with interactions (b) random–intercept models. Each panel shows a histogram of the estimated intercepts over-laid with its kernel-density curve.

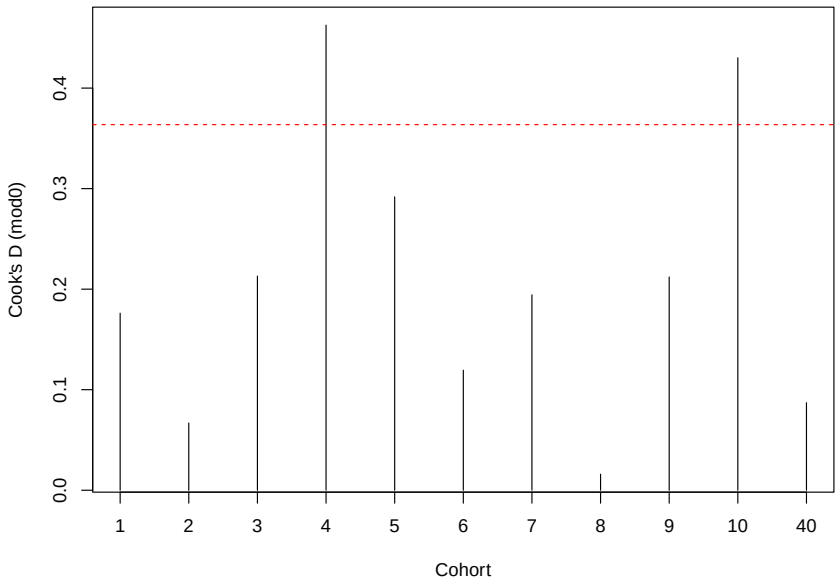


Figure 19: Cluster-level influence measures: Cook's distance per cohort in the baseline mixed-effects model. The horizontal dashed line at (4/n) indicates the conventional influence threshold.

B Additional tables

Table 10: For information only: full set of biomarkers included in the HBM4EU Children dataset

Variable	Biomarker (full name)	Characteristics
mbzp	***mono-benzyl phthalate***	***Primary metabolite of benzyl butyl phthalate (BBzP)***
mibp	mono-isobutyl phthalate	Primary metabolite of di-isobutyl phthalate (DiBP)
mnbp	mono-*n*-butyl phthalate	Primary metabolite of di-*n*-butyl phthalate (DnBP)
mehp	mono-2-ethylhexyl phthalate	Primary metabolite of di-2-ethylhexyl phthalate (DEHP)
oh-mehp	mono-(2-ethyl-5-hydroxyhexyl) phthalate	Secondary metabolite of DEHP
oxo-mehp	mono-(2-ethyl-5-oxohexyl) phthalate	Secondary metabolite of DEHP
cx-mepp	mono-(2-ethyl-5-carboxypentyl) phthalate	Secondary metabolite of DEHP
mcoch	mono-(2-ethyl-5-carboxyhexyl) phthalate	Secondary metabolite of DEHP
mep	monoethyl phthalate	Primary metabolite of diethyl phthalate (DEP)
mhnp	mono-(4-hydroxy-nonyl) phthalate	Secondary metabolite of di-*n*-octyl phthalate (DnOP)
mcop	mono-cyclohexyl phthalate	Primary metabolite of di-cyclohexyl phthalate (DCHP)
oh-midp	mono-(2-ethyl-5-hydroxyhexyl) isodecyl phthalate	Secondary metabolite of di-isodecyl phthalate (DIDP)
cx-midp	mono-(2-ethyl-5-carboxypentyl) isodecyl phthalate	Secondary metabolite of DIDP
mhnch	mono-(3-hydroxy-*n*-hexyl) phthalate	Secondary metabolite of di-*n*-hexyl phthalate (DnHP)
sum-midp	oh-midp: mono-(2-ethyl-5-hydroxyhexyl) isodecyl phthalate cx-midp: mono-(2-ethyl-5-carboxypentyl) isodecyl phthalate	sum of 2 secondary DIDP metabolites
sum-minch	oh-minch: mono-(2-ethyl-5-hydroxyhexyl) cyclohexyl phthalate cx-minch: mono-(2-ethyl-5-carboxypentyl) cyclohexyl phthalate	sum of 2 secondary DINCH metabolites
sum-minp	mhnp: mono-(4-hydroxy-nonyl) phthalate mcop: mono-cyclohexyl phthalate	sum of 2 metabolites: (mhnp secondary from DnOP; mcop primary from DCHP)
sum-ohcxmehp	oh-mehp: mono-(2-ethyl-5-hydroxyhexyl) phthalate cx-mepp: mono-(2-ethyl-5-carboxypentyl) phthalate	sum of 2 secondary DEHP metabolites

sum-ohoxocxmehp	oxo-mehp: mono (2-ethyl-5-oxohexyl) phthalate oh-mehp: mono (2-ethyl-5-hydroxyhexyl) phthalate mehp: mono-2-ethylhexyl phthalate cx-mepp: mono (2-ethyl-5-carboxypentyl) phthalate	sum of 1 primary (MEHP) and 3 secondary DEHP metabolites
sumohoxomehp	oxo-mehp: mono-(2-ethyl-5-oxohexyl) phthalate oh-mehp: mono-(2-ethyl-5-hydroxyhexyl) phthalate	sum of 2 secondary DEHP metabolites
sumohoxomehptwo	oxo-mehp: mono-(2-ethyl-5-oxohexyl) phthalate oh-mehp: mono-(2-ethyl-5-hydroxyhexyl) phthalate mehp: mono-2-ethylhexyl phthalate	sum of 1 primary (mehp) and 2 secondary (oh-mehp, oxo-mehp) DEHP metabolites
sum-eightphthal	mbzp: mono-benzyl phthalate mibp: mono-isobutyl phthalate mnbp: mono-*n*-butyl phthalate mep: monoethyl phthalate mehp: mono-2-ethylhexyl phthalate oh-mehp: mono-(2-ethyl-5-hydroxyhexyl) phthalate oxo-mehp: mono-(2-ethyl-5-oxohexyl) phthalate cx-mepp: mono-(2-ethyl-5-carboxypentyl) phthalate	sum of 8 phthalate metabolites: • 4 primary (mbzp, mibp, mnbp, mep) • 4 DEHP metabolites (mehp [primary], oh-mehp, oxo-mehp, cx-mepp [3 secondary])
sumlmwphthal	mibp: mono-isobutyl phthalate mnbp: mono-n-butyl phthalate mep: monoethyl phthalate	sum of 3 primary low-molecular-weight metabolites
sumhmwphthal	mbzp: mono-benzyl phthalate mehp: mono-2-ethylhexyl phthalate oh-mehp: mono-(2-ethyl-5-hydroxyhexyl) phthalate oxo-mehp: mono-(2-ethyl-5-oxohexyl) phthalate cx-mepp: mono-(2-ethyl-5-carboxypentyl) phthalate	sum of 5 high-molecular-weight metabolites: • mbzp (primary from (BBzP) • mehp (primary), oh-mehp, oxo-mehp, cx-mepp (DEHP metabolites; 3 secondary)
`bdcippms`	bis(1,3-dichloro-2-propyl) phosphate monoester	primary metabolite of tris(1,3-dichloro-2-propyl) phosphate (TDCIPP), an organophosphate flame retardant
`dphpms`	diphenyl phosphate monoester	primary metabolite of triphenyl phosphate (TPHP), an organophosphate flame retardant

C R-code

```
----Methodology----
----Methodology 3.1 Part 1----
#using Eurostat data with both males and females, 6-12 yrs, only EU27
#loading DOI: 10.2908/demo_pjan
eurostat_ages_MF_EU <- "/Users/vivif/R/demo_pjan__males AND females_all ages._ONLY EU.xlsx"
#loading dataset from HBM4EU
data_HBMC <- read.xlsx("C:/Users/vivif/R/CHILDREN_PEH_2023_19c-UH-2024-05-06-16-07.xlsx", sheet = 1)

# Importing the raw data (for ages: 6-12)
sheets <- 2:8
ages   <- 6:12
age_dfs <- lapply(sheets, function(i) {
  read_excel(eurostat_ages_MF_EU, sheet = i)})

# Annotating each data frame with its age
age_dfs <- Map(function(df, a) mutate(df, age = a), age_dfs, ages)

# Combining all data into one data frame
combined_data_MF_ages_EU <- bind_rows(age_dfs)

# Aggregating the population across all age groups (6-12) for each country
aggregated_data_MF_ages_EU <- combined_data_MF_ages_EU %>%
  group_by(Countries) %>%
  summarise(total_population_6_12_MF_EU = sum(Population, na.rm = TRUE))

# Creating region mapping for only EU-27 using a compact list + stack approach
regions <- list(
  West  = c("Austria","Belgium","France","Germany","Luxembourg","Netherlands"),
  North = c("Denmark","Estonia","Finland","Ireland","Latvia","Lithuania","Sweden"),
  East  = c("Bulgaria","Czechia","Hungary","Poland","Romania","Slovakia"),
  South = c("Croatia","Cyprus","Greece","Italy","Malta","Portugal","Slovenia","Spain")

region_mapping_EU27 <- stack(regions) %>%
  rename( Countries = values,
          region    = ind )
# Merging aggregated data with region mapping
combined_MF_ages_EU27 <- left_join(
  aggregated_data_MF_ages_EU,
  region_mapping_EU27,
  by = "Countries")

#group by region and sum the population
aggregated_by_region_EU <- combined_MF_ages_EU27 %>%
  group_by(region) %>%
  summarise(total_population = sum(total_population_6_12_MF_EU, na.rm = TRUE))

# age{region distribution - all ages from 6 to 12 yrs
age_region_dist <- combined_data_MF_ages_EU %>%
  left_join(region_mapping_EU27, by = "Countries") %>%
  group_by(region, age) %>%
  summarise(count = sum(Population, na.rm = TRUE), .groups = "drop") %>%
  group_by(region) %>%
```

```
  mutate(prop = count / sum(count)) %>%
  ungroup()

----Methodology 3.1 Part 2----
#using Eurostat data to identify DEGURBA distributions, only EU27
#loading 'Distribution of population by DEGURBA, dwelling type (total) and income group(total)'
#DOI:10.2908/ilc_lvho01, % per country, per DEGURBA

# DEGURBA percentages by class
degurba_percent <- list(
  Urban = "1 Cities",
  `Towns & Suburbs` = "2 Towns& suburbs",
  Rural = "3 Rural") %>%
  purrr::imap_dfr(~ read_excel("degurba_population percentage by country_EU27.xlsx", sheet = .x) %>%
  mutate(DEGURBA = .y))

# Total pop and children EU27 only
overall_pop <- read_excel("Total population per EU country.xlsx", sheet = 2)
children_pop <- read_excel("Total population per EU country_for 6-12 yrs old.xlsx")

# Merging population data, calculating ratio
pop_combined <- overall_pop %>%
  inner_join(children_pop, by = "Countries", suffix = c("_total", "_children")) %>%
  mutate(ratio = Population_children / Population_total)

# Merging with DEGURBA
combined_estimates <- degurba_percent %>%
  left_join(pop_combined, by = "Countries") %>%
  mutate(estimated_children = Population_total * (`Population perc.` / 100) * ratio)

# Adding region info
combined_estimates_f <- combined_estimates %>%
  left_join(region_mapping_EU27, by = "Countries")

# Aggregating
aggregated_by_region_deg <- combined_estimates_f %>%
  group_by(region, DEGURBA) %>%
  summarise(total_population = sum(estimated_children, na.rm = TRUE), .groups = "drop")

----Methodology 3.1 Part 3----
# Reading education dataset from Eurostat
edu_data <- read_excel("eurostat_edu level_experimental_household.xlsx",
  sheet = 2) %>%
  rename(Country= 1, TertiaryExclShort  = `Tertiary education excluding short cycle`)

# Summarizing into Low / Medium / High ISCED groups;summing rowwise into 3 groups:
# - Low Education: ISCED levels 0--2 (Early childhood, Primary, Lower secondary)
# - Medium Education: ISCED levels 3--4 (Upper secondary, Post-secondary non-tertiary)
# - High Education: ISCED $\ge5$5 (Short-cycle tertiary, Tertiary excluding short-cycle)
edu_data_summary <- edu_data %>%
  rowwise() %>%
  mutate(
    LowEducation = sum(c_across(c(
      `Early childhood education`,
```

```
      `Primary education`,
      `Lower secondary education`
    )), na.rm = TRUE),
    MediumEducation = sum(c_across(c(
      `Upper secondary education`,
      `Post-secondary non-tertiary education`
    )), na.rm = TRUE),
    HighEducation = sum(c_across(c(
      `Short-cycle tertiary education`,
      TertiaryExclShort)), na.rm = TRUE)) %>% ungroup()

library(dplyr)
# Normalising education shares to sum to 100 per country
normalized_edu <- edu_data_summary %>%
  mutate(
    Total               = LowEducation + MediumEducation + HighEducation,
    LowEducation_norm   = (LowEducation   / Total) * 100,
    MediumEducation_norm= (MediumEducation/ Total) * 100,
    HighEducation_norm  = (HighEducation  / Total) * 100
  ) %>%
  dplyr::select(
    Country,
    LowEducation_norm,
    MediumEducation_norm,
    HighEducation_norm)

# joining total child population per EU27 country, with normalized education shares
edu_pop <- aggregated_data_MF_ages_EU %>%
  rename(Country = Countries) %>%
left_join(normalized_edu, by = "Country")
# adding region info for each country
edu_pop_reg <- edu_pop %>%
  left_join(region_mapping_EU27, by = c("Country" = "Countries")) %>%
  # re-factoring 'region' so that its levels are North|South|West|East
  mutate(region = factor(region, levels = c("North","South","West","East")))

edu_pop_reg_class <- edu_pop_reg %>%
  mutate(
    LowCount  = total_population_6_12_MF_EU * LowEducation_norm   / 100,
    MedCount  = total_population_6_12_MF_EU * MediumEducation_norm/ 100,
    HighCount = total_population_6_12_MF_EU * HighEducation_norm  / 100)

aggregated_by_region_edu <- edu_pop_reg_class %>%
  group_by(region) %>%
  summarise(
    Low    = sum(LowCount,  na.rm = TRUE),
    Medium = sum(MedCount,  na.rm = TRUE),
    High   = sum(HighCount, na.rm = TRUE),
    .groups = "drop")
```

```
----Methodology 3.2----
## ---------------- External EU-27 margins ----------------
# 1. Region external proportions (Eurostat; children 6 to 12 yrs)
external_region <- aggregated_by_region_EU %>%
  mutate(
    region = factor(region,
                        levels = c("North","South","West","East")),
    region_prop = total_population / sum(total_population)) %>%
  arrange(region)
print(external_region)
#table with raw counts, then normalized to proportions (region_prop)
#North ~8.46%, South ~27.4%, West ~44%, East ~20.2%

# 2. Sex external proportions (Eurostat; children 6{12 yrs)
# Function to sum ages 6{12 in a given Excel file
sum_ages_6_12 <- function(path) {
  sheets <- 2:8
  dfs <- lapply(sheets, function(i) {
    read_excel(path, sheet = i) %>%
      dplyr::select(Population)
  })
  bind_rows(dfs) %>%
    summarise(total = sum(Population, na.rm = TRUE)) %>%
    pull(total)}
# File paths
male_file   <- "demo_pjan__males across all ages.xlsx"
female_file <- "demo_pjan__females across all ages.xlsx"

# total counts for each sex
male_tot   <- sum_ages_6_12(male_file)
female_tot <- sum_ages_6_12(female_file)

# external sex margin (M as reference level)
external_sex <- tibble(
  sex   = factor(c("M", "F"), levels = c("M", "F")),
  count = c(male_tot, female_tot)
) %>%
  mutate(
    sex_prop = count / sum(count))
# Approximately: M ~51.4%, F ~48.6%

# 3. Sampling season: using numeric values (1, 2, 3, 4) representing Spring, Summer, Autumn, Winter
external_season <- tibble(
samplingseason = factor(1:4,
                         levels = 1:4,
                         labels = c("Spring","Summer","Autumn","Winter")),
season_prop = rep(0.25, 4))
#assuming equal 25% in spring/summer/autumn/winter

# 4. DEGURBA external proportions across the EU (based on Eurostat data)
external_degurba <- aggregated_by_region_deg %>%
  group_by(region) %>%
  mutate(region_degurba_prop = total_population / sum(total_population)) %>%
  ungroup() %>%
```

```
  left_join(
    # grab just the two columns by name
    external_region[ , c("region", "region_prop") ],
    by = "region"
  ) %>%
  mutate(weighted = region_degurba_prop * region_prop) %>%
  group_by(DEGURBA) %>%
  summarise(
    degurba_prop = sum(weighted, na.rm = TRUE),
    .groups      = "drop"
  ) %>%
  mutate(
    DEGURBA = factor(
      DEGURBA,
      levels = c("Urban", "Towns & Suburbs", "Rural")))
# approximate proportions: Urban ~0.385, Towns ~0.333, Rural ~0.282

# 5. ISCED external proportions:
external_isced <- aggregated_by_region_edu %>%
    # a) Pivoting to long form, one row per region{ISCED level
  pivot_longer(
    cols      = c(Low, Medium, High),
    names_to  = "isced_hh",
    values_to = "count") %>%
  # b) Computing each region's within-region ISCED share
  group_by(region) %>%
  mutate(
    region_total      = sum(count),
    region_isced_prop = count / region_total) %>%
  ungroup() %>%
   # c) Attaching the EU-wide region_prop for each region
  left_join(
    # force dplyr::select() to avoid masking issues
    external_region %>% dplyr::select(region, region_prop),
    by = "region"
  ) %>%
  # d) Multiplying the within-region ISCED share by that region's EU weight
  mutate(
    weighted = region_isced_prop * region_prop) %>%
  # e) Summing across all regions, grouping by ISCED level
  group_by(isced_hh) %>%
  summarise(
    weighted_total = sum(weighted, na.rm = TRUE),
    .groups        = "drop") %>%
   # f) Converting to true EU-wide proportions
  mutate(
    isced_prop = weighted_total / sum(weighted_total))
# approximately: Low ~26%, Medium ~41.9%, High ~32.1%
#coercing isced_hh to a factor
external_isced <- external_isced %>%
  mutate(
    isced_hh = factor(isced_hh, levels = c("Low","Medium","High")))

## Building the reference grid "average EU child"
```

```
# Creating a cross-product combination of all external margins
ref_grid_base <- tidyr::crossing(
  external_region,
  external_sex,
  external_season,
  external_degurba,
  external_isced) %>%
  mutate(
    ageyears = 9,
    overall_weight = region_prop * sex_prop * season_prop * degurba_prop * isced_prop) %>%
  rename(degurba = DEGURBA) %>%
  dplyr::select(region, sex, ageyears, samplingseason, degurba, isced_hh, overall_weight)

# Check that weights sum to 1:
sum(ref_grid_base$overall_weight)

#PLOT for weights distribution - reference grid
## quick EDA of overall_weight
# numeric summary
summ_w <- ref_grid_base %>%
  summarise(min   = min(overall_weight),
            p1    = quantile(overall_weight, 0.01),
            p5    = quantile(overall_weight, 0.05),
            median= median(overall_weight),
            mean  = mean(overall_weight), # approx 1/288 = 0.00347
            p95   = quantile(overall_weight, 0.95),
            p99   = quantile(overall_weight, 0.99),
            max   = max(overall_weight),
            N_heavy = sum(overall_weight > 0.01))  # cells > 1 %
print(round(summ_w, 5))

# histogram on a log10 scale
ggplot(ref_grid_base, aes(x = overall_weight)) +
  geom_histogram(
    binwidth = 0.1,
    colour   = "black",
    fill     = "grey70") +
  scale_x_log10(
    labels = label_scientific(digits = 1)) +
  labs(
    x = "Cell weight (log10 scale)",
    y = "Frequency") +
  theme_minimal(base_size = 11) +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    panel.border     = element_blank())

# where do the heaviest cells live?
top_cells <- ref_grid_base %>%
  arrange(desc(overall_weight)) %>%
  slice_head(n = 10) # top 10 largest weights
print(top_cells, n = 10)
```

```
#----Methodology 3.3----
#----Methodology 3.3.1----
##Prediction & Overall Aggregation for simple linear model

#checking for any values beyong LOD and LOQ limits
any(data_HBMC$mbzp_impcrtlog %in% c(-1, -2, -3))

#Study data: filtering, recodimg, *centering for age*
dataHBMC_filtered <- data_HBMC %>%
  filter(between(ageyears, 6, 12), region %in% 1:4) %>%      # keep 6{12 yrs, EU-27 regions
  mutate(
    region         = factor(region,     1:4, c("North","South","West","East")),
    degurba        = factor(degurba,    1:3, c("Urban","Towns & Suburbs","Rural")),
    isced_hh       = factor(isced_hh,   1:3, c("Low","Medium","High")),
    sex            = factor(sex,        c("M","F")),
    samplingseason = factor(samplingseason, 1:4,
                            c("Spring","Summer","Autumn","Winter")),
    age_c          = ageyears - 9)

# Fitting a basic linear model (no interactions) to the filtered study data, estimating coefficients from the study
lm_base <- lm(
  mbzp_impcrtlog ~ region + sex + age_c +
    degurba + isced_hh + samplingseason,
  data = dataHBMC_filtered)
summary(lm_base)
confint(lm_base)

# predicting on the 288-cell reference grid (adding age_c = 0)
pred_grid <- ref_grid_base %>%
  mutate(age_c = 0) %>%       # grid represents a 9-years-old midpoint
  bind_cols(predict(lm_base, newdata = ., se.fit = TRUE)) %>%
  rename(predicted = fit, se_pred = se.fit)

# EU-wide mean and standard error (weights already sum to 1)
eu_mean <- with(pred_grid, sum(predicted * overall_weight))
eu_se   <- with(pred_grid, sqrt(sum((overall_weight)^2 * se_pred^2)))

# geometric mean on the original scale
eu_geo_mean <- exp(eu_mean)

# 95 % confidence interval on the original scale
ci_geo <- exp(eu_mean + c(-1.96, 1.96) * eu_se)

cat("--- BASE LINEAR MODEL (centred age) ---\n",
    sprintf("EU-wide mean ln-MBzP      : %.4f  (SE = %.4f)\n", eu_mean, eu_se),
    sprintf("EU geometric mean         : %.4f µg g-creatinine\n", eu_geo_mean),
    sprintf("95%% CI (geo-scale)        : [%.4f, %.4f]\n", ci_geo[1], ci_geo[2]))

#----Methodology 3.3.2----
library(lmtest)      # lrtest
library(broom)       # tidy / confint
library(dplyr)
library(tibble)
```

```
library(emmeans)     # emmeans, emtrends
library(purrr)

# lm_base includes the centred age term age_c = ageyears - 9
# Fitting one-by-one 2-way interaction models
int_terms <- c(
  "region:samplingseason",   # region × season
  "region:age_c",            # region × age (centred)
  "sex:age_c",               # sex × age (centred)
  "region:degurba",          # region × DEGURBA
  "samplingseason:degurba"   # season × DEGURBA)

int_models <- setNames(
  lapply(int_terms, function(trm)
    update(lm_base, paste(". ~ . +", trm))),
  int_terms)
all_models <- c(Base = list(lm_base), int_models)

#AIC & LR comparisons
compare_tbl <- tibble(
  Model = names(all_models),
  df    = sapply(all_models, function(m) attr(logLik(m), "df")),
  AIC   = sapply(all_models, AIC)) %>%
  mutate(
    DeltaAIC = AIC - AIC[Model == "Base"],
    LR_df    = c(NA, sapply(int_models, function(m) lrtest(lm_base, m)$Df[2])),
    LR_chi   = c(NA, sapply(int_models, function(m) lrtest(lm_base, m)$Chisq[2])),
    p_value  = c(NA, sapply(int_models, function(m) lrtest(lm_base, m)$`Pr(>Chisq)`[2])))

cat("\n### AIC and likelihood-ratio comparison\n")
print(compare_tbl, digits = 3)

# Coefficient tables (all models), with 95% CIs
coef_df <- function(fit, level = 0.95) {
  summ <- summary(fit)$coefficients
  ci   <- confint(fit, level = level)
  tibble(
    term      = rownames(summ),
    Estimate  = summ[, "Estimate"],
    Std_Error = summ[, "Std. Error"],
    t_value   = summ[, "t value"],
    p_value   = summ[, "Pr(>|t|)"],
    CI_lower  = ci[, 1],
    CI_upper  = ci[, 2]
  )}

for (nm in names(all_models)) {
  cat("\n### Coefficients for", nm, "\n")
  print(coef_df(all_models[[nm]]), digits = 4)}

## Biologically interpretable contrasts via emmeans (for all tested interactions)
emm_to_df <- function(obj, by_var) {
  out <- as.data.frame(summary(obj, infer = TRUE))
  names(out) <- tolower(names(out))
```

```
  est_col <- grep("^(emmean|estimate|.*\\.trend)$", names(out), value = TRUE)[1]
  se_col  <- grep("^(se$|se\\.trend$)",            names(out), value = TRUE)[1]

  out %>%
    dplyr::rename(estimate = all_of(est_col),
                  se       = all_of(se_col)) %>%
    dplyr::mutate(across(all_of(by_var), as.character), .before = 1) %>%
    dplyr::select(all_of(by_var), estimate, se, df, lower.cl, upper.cl, p.value)}

contr_tables <- list()

# 1. Winter{Spring by region (region × season)
emm_rs <- emmeans(int_models[["region:samplingseason"]],
                  ~ samplingseason | region)
ws_con <- contrast(emm_rs,
                   list("Winter{Spring" = c(-1, 0, 0, 1)),
                   adjust = "none")
contr_tables[["region:samplingseason"]] <-
  emm_to_df(ws_con, "region") %>%
  mutate(contrast = "Winter{Spring", .before = 1)

# 2. Age slope by region (region × age_c)
sl_ra <- emtrends(int_models[["region:age_c"]],
                  ~ region, var = "age_c")
contr_tables[["region:age_c"]] <-
  emm_to_df(sl_ra, "region") %>%
  mutate(measure = "Age slope (Δ log / yr)", .before = 1)

# 3. Age slope by sex (sex × age_c)
sl_sa <- emtrends(int_models[["sex:age_c"]],
                  ~ sex, var = "age_c")
contr_tables[["sex:age_c"]] <-
  emm_to_df(sl_sa, "sex") %>%
  mutate(measure = "Age slope (Δ log / yr)", .before = 1)

# 4. Rural{Urban by region (region × DEGURBA)
emm_rd <- emmeans(int_models[["region:degurba"]],
                  ~ degurba | region)
ru_con <- contrast(emm_rd,
                   list("Rural{Urban" = c(-1, 0, 1)),
                   adjust = "none")
contr_tables[["region:degurba"]] <-
  emm_to_df(ru_con, "region") %>%
  mutate(contrast = "Rural{Urban", .before = 1)

# 5. Rural{Urban by season (season × DEGURBA)
emm_sd <- emmeans(int_models[["samplingseason:degurba"]],
                  ~ degurba | samplingseason)
ru_seas <- contrast(emm_sd,
                    list("Rural{Urban" = c(-1, 0, 1)),
                    adjust = "none")
contr_tables[["samplingseason:degurba"]] <-
  emm_to_df(ru_seas, "samplingseason") %>%
```

```
  mutate(contrast = "Rural{Urban", .before = 1)

for (nm in names(contr_tables)) {
  cat("\n### Contrast results for", nm, "\n")
  print(contr_tables[[nm]], digits = 4)}

## 2-way interactions candidate
mod_two_int <- update(
  lm_base,
  . ~ . + region:samplingseason + region:degurba)
#AIC and LR test
delta_aic_two <- AIC(mod_two_int) - AIC(lm_base)
cat("\n### Combined model: region:samplingseason + region:degurba\n")
cat(sprintf("AIC  = %.1f   (ΔAIC = %.1f relative to main effects)\n",
            AIC(mod_two_int), delta_aic_two))

lr_two_int <- lrtest(lm_base, mod_two_int)
print(lr_two_int)

#  CONTRASTS FOR THE COMBINED MODEL  (region:season + region:degurba)
# ------------------------------------------------------------
## Helper: convert an emmeans / contrast summary to tidy tibble
emm_to_df <- function(obj, by_var) {
  out <- as.data.frame(summary(obj, infer = TRUE))
  names(out) <- tolower(names(out))

  est_col <- grep("^(emmean|estimate|.*\\.trend)$", names(out), value = TRUE)[1]
  se_col  <- grep("^(se$|se\\.trend$)",             names(out), value = TRUE)[1]

  out %>%
    dplyr::rename(estimate = all_of(est_col),
                  se       = all_of(se_col)) %>%
    dplyr::mutate(across(all_of(by_var), as.character), .before = 1) %>%
    dplyr::select(all_of(by_var), estimate, se, df,
                  lower.cl, upper.cl, p.value)}

# a) Winter { Spring contrast by REGION  (degurba = "Urban")
emm_rs <- emmeans(
  mod_two_int,
  ~ samplingseason | region,
  at = list(degurba = "Urban")      # hold other factor at reference
)
ws_con <- contrast(
  emm_rs,
  list("Winter{Spring" = c(-1, 0, 0, 1)),  # Spring, Summer, Autumn, Winter
  adjust = "none"
) %>%
  emm_to_df("region") %>%
  mutate(contrast = "Winter{Spring", .before = 1)

# b) Rural { Urban contrast by REGION  (samplingseason = "Spring")
emm_rd <- emmeans(
  mod_two_int,
  ~ degurba | region,
```

```
  at = list(samplingseason = "Spring")      # season at reference)

ru_con <- contrast(
  emm_rd,
  list("Rural{Urban" = c(-1, 0, 1)),       # Urban, Towns&Suburbs, Rural
  adjust = "none") %>%
  emm_to_df("region") %>%
  mutate(contrast = "Rural{Urban", .before = 1)

# c) Combining and keeping only significant rows (p < .05)
key_contrasts <- bind_rows(ws_con, ru_con) %>%
  mutate(across(where(is.numeric), round, 3)) %>%
  filter(p.value < 0.05)

# EU-standardised means: base + selected interactions
# ------------------------------------------------
grid_int <- ref_grid_base %>% mutate(age_c = 0)

eu_summary <- function(fit, grid = grid_int) {
  pr <- predict(fit, newdata = grid, se.fit = TRUE)
  wt <- grid$overall_weight
  tibble(
    mean = sum(pr$fit * wt),
    se   = sqrt(sum((wt^2) * pr$se.fit^2)))}

eval_models <- list(
  Base                = lm_base,
  INT_region_season   = update(lm_base, . ~ . + region:samplingseason),
  INT_region_degurba  = update(lm_base, . ~ . + region:degurba),
  INT_both            = update(lm_base, . ~ . + region:samplingseason + region:degurba))

eu_results_int <- imap_dfr(
  eval_models,
  ~ eu_summary(.x) %>% mutate(model = .y),
  .id = NULL
) %>%
  relocate(model) %>%
  mutate(
    geo_mean = exp(mean),
    geo_lwr  = exp(mean - 1.96 * se),
    geo_upr  = exp(mean + 1.96 * se))

cat("\n### EU-standardised results (log and original scale)\n")
print(eu_results_int, digits = 4)

#----Methodology 3.3.3----
#-----MIXED MODELS-----
library(lme4)          # lmer(), VarCorr()
library(MASS)          # mvrnorm()
library(dplyr)
library(broom)         # tidy(), glance()
library(car)           # vif()
library(influence.ME)  # influence diagnostics
library(purrr)         # imap_dfr (for sensitivity block)
```

```
# Objects already in memory
#  * dataHBMC_filtered   { HBM4EU children, ages 6{12,  recoded factors
#  * ref_grid_base       { 288-cell EU-27 reference grid (+ overall_weight))

# Pre-processing
dataHBMC_filtered <- dataHBMC_filtered %>%
  mutate(
    cohort = factor(cohort),            # 11 study IDs
    age_c  = ageyears - 9)              # centred age (mid-point = 9 y)

ref_grid_base_mixed <- ref_grid_base %>%
  mutate(age_c = 0)                     # grid represents a 9-year-old child

w_mixed <- ref_grid_base_mixed$overall_weight  # numeric(288); sum = 1

#----convenience helpers----
#Helper: extract random-intercept variance (σ²_u)
get_sigma_u2 <- function(fit, grp = "cohort") {
  as.numeric(VarCorr(fit)[[grp]][1, 1])  # first (and only) element of 1×1 vcov}
}
#Helper: intraclass-correlation coefficient (ICC)
icc <- function(fit) {
  sig_u2 <- get_sigma_u2(fit)
  sig_e2 <- sigma(fit)^2
  sig_u2 / (sig_u2 + sig_e2)}

# Helper: Delta-method + Monte-Carlo SE on a design grid ---------------
# Returns a list with means & SEs (Delta, MC-fixed, MC-full)
#   fit     : lmer() object
#   grid    : data.frame with same factor levels as fit
#   weights : numeric vector, length = nrow(grid), summing to 1
mixed_SE <- function(fit, grid, weights, n_sims = 5000, seed = 2025) {
  stopifnot(isTRUE(all.equal(sum(weights), 1, tol = 1e-12)))

# fixed-effects component on the grid
X <- model.matrix(delete.response(terms(fit, fixed.only = TRUE)), grid)
b <- fixef(fit)
V <- vcov(fit)

xb     <- as.numeric(X %*% b)             # linear predictor (β part)
se_xb  <- sqrt(rowSums((X %*% V) * X))    # sqrt(diag(X V XT))
sig_u2 <- get_sigma_u2(fit)               # random-intercept variance

# Delta-method
mu_delta <- sum(xb * weights)
se_delta <- sqrt(sum(weights^2 * (se_xb^2 + sig_u2)))

# Monte-Carlo
set.seed(seed)
beta_draws <- MASS::mvrnorm(n_sims, mu = b, Sigma = V)
xb_mat <- X %*% t(beta_draws)
mu_fix <- colSums(xb_mat * weights)
se_fix <- sd(mu_fix)
```

```
u_draws <- rnorm(n_sims, sd = sqrt(sig_u2))
xb_full <- sweep(xb_mat, 2, u_draws, "+")
mu_full <- colSums(xb_full * weights)
se_full <- sd(mu_full)

list(mu_delta = mu_delta, se_delta = se_delta,
       mu_fix = mean(mu_fix), se_fix = se_fix,
       mu_full = mean(mu_full), se_full = se_full)}

# Helper: print one result block
print_mixed_res <- function(res, label) {
  cat("\n|", label, "|\n", sep = "")
  with(res, {
    cat(sprintf("Δ-method : mean = %.4f   SE = %.4f\n", mu_delta, se_delta))
    cat(sprintf("MC fixed : mean = %.4f   SE = %.4f\n", mu_fix,   se_fix))
    cat(sprintf("MC full  : mean = %.4f   SE = %.4f\n", mu_full,  se_full))})}

# Baseline mixed model
mod0 <- lmer(mbzp_impcrtlog ~ region + sex + age_c +
               degurba + isced_hh + samplingseason +
               (1 | cohort),
             data = dataHBMC_filtered, REML = FALSE)
cat(sprintf("Baseline ICC = %.3f\n", icc(mod0)))

res_mod0 <- mixed_SE(mod0, ref_grid_base_mixed, w_mixed)
print_mixed_res(res_mod0, "Baseline mixed model")
cat(sprintf("\nΔ-GM = %.4f  (95%% CI %.4f { %.4f)\n",
            exp(res_mod0$mu_delta),
            exp(res_mod0$mu_delta - 1.96 * res_mod0$se_delta),
            exp(res_mod0$mu_delta + 1.96 * res_mod0$se_delta)))

## Diagnostics mod0
fitted0 <- fitted(mod0); stdres0 <- resid(mod0) / sigma(mod0)
plot(fitted0, stdres0, xlab = "Fitted (mod0)", ylab = "Std. residuals"); abline(h=0,lty=2)
qqnorm(stdres0); qqline(stdres0, col="red")
blups0 <- ranef(mod0)$cohort[, 1]
hist(
  blups0,
  breaks = "FD",
  prob   = TRUE,
  xlab   = "Random-intercept BLUPs (mod0)",
  main   = NULL)
lines(density(blups0), lwd = 1.5)

cd0 <- cooks.distance(influence(mod0, group="cohort", prune.fixed = TRUE))
plot(cd0, type="h", xaxt="n", xlab="Cohort",
ylab="Cook's D (mod0)"); axis(1, at=seq_along(cd0),
labels=rownames(cd0)); abline(h=4/length(cd0), col="red", lty=2)

# Leave-one-cohort-out sensitivity
# Helper: EU-standardised mean & SE for a fitted mode
eu_mean_se <- function(fit, grid, weights) {
  pr <- predict(fit, newdata = grid, re.form = NA, se.fit = TRUE)
```

```
  se_tot <- sqrt(pr$se.fit^2 + get_sigma_u2(fit))
  tibble(mean = sum(pr$fit * weights),
         se   = sqrt(sum(weights^2 * se_tot^2)))}

# Build a grid compatible with the data in `dat`
make_grid <- function(dat) {
  ref_grid_base_mixed %>%
    filter(region         %in% unique(dat$region),
           samplingseason %in% unique(dat$samplingseason),
           degurba        %in% unique(dat$degurba),
           isced_hh       %in% unique(dat$isced_hh),
           sex            %in% unique(dat$sex))}

# Cohort-removal scenarios
scenarios <- list(
  keep_all  = dataHBMC_filtered,
  drop_4    = filter(dataHBMC_filtered,  cohort != "4"),
  drop_10   = filter(dataHBMC_filtered,  cohort != "10"),
  drop_4_10 = filter(dataHBMC_filtered, !cohort %in% c("4", "10")))

sensitivity_tbl <- purrr::imap_dfr(scenarios, function(dat, label) {
  grid <- make_grid(dat)
  wt   <- grid$overall_weight / sum(grid$overall_weight)  # renormalise Σw = 1

  fit <- lmer(mbzp_impcrtlog ~ region + sex + age_c +
                degurba + isced_hh + samplingseason +
                (1 | cohort),
              data = dat, REML = FALSE)

  eu  <- eu_mean_se(fit, grid, wt)
  mutate(eu, scenario = label, .before = 1)})

print(sensitivity_tbl, digits = 4)

# ----Interaction model----
mod_int <- lmer(mbzp_impcrtlog ~ region * samplingseason +
region * degurba + sex + age_c + isced_hh + (1 | cohort),
                data = dataHBMC_filtered, REML = FALSE)
cat(sprintf("Interaction ICC = %.3f\n", icc(mod_int)))
print(anova(mod0, mod_int))

res_int <- mixed_SE(mod_int, ref_grid_base_mixed, w_mixed)
print_mixed_res(res_int, "Interaction model")
cat(sprintf("\nΔ-GM = %.4f  (95%% CI %.4f { %.4f)\n",
            exp(res_int$mu_delta),
            exp(res_int$mu_delta - 1.96 * res_int$se_delta),
            exp(res_int$mu_delta + 1.96 * res_int$se_delta)))

## Interaction-model diagnostics
fittedI <- fitted(mod_int); stdresI <- resid(mod_int) / sigma(mod_int)
plot(fittedI, stdresI, xlab="Fitted (mod_int)", ylab="Std. residuals"); abline(h=0,lty=2)
qqnorm(stdresI); qqline(stdresI, col="red")
blupsI <- ranef(mod_int)$cohort[, 1]
hist(
```

```
  blupsI,
  breaks = "FD",
  prob   = TRUE,
  xlab   = "Random-intercept BLUPs (mod_int)",
  main   = NULL        )
lines(density(blupsI), lwd = 1.5)

## VIF table
lm_fixed_int <- lm(mbzp_impcrtlog ~ region * samplingseason +
region * degurba + sex + age_c + isced_hh, data = dataHBMC_filtered)
print(vif(lm_fixed_int))

# Cook's distance plot for the interaction model omitted
# (rank deficiency in several leave-one-cohort fits made the metric unstable)

# ----- diagnostics of OLS for comparison ------
# OLS baseline (Model 1) and full interaction (Model 4)
lm1 <- lm(mbzp_impcrtlog ~ region + sex + age_c +
            degurba + isced_hh + samplingseason,
          data = dataHBMC_filtered)

lm4 <- lm(mbzp_impcrtlog ~ region * samplingseason +
            region * degurba + sex + age_c + isced_hh,
          data = dataHBMC_filtered)

# Quick base-R residual diagnostics, without observation numbers
par(mfrow = c(2, 2))
plot(lm1, which = 1, id.n = 0, main = "OLS Model 1: Residuals vs Fitted")
plot(lm1, which = 2, id.n = 0, main = "OLS Model 1: QQ-plot")
plot(lm4, which = 1, id.n = 0, main = "OLS Model 4: Residuals vs Fitted")
plot(lm4, which = 2, id.n = 0, main = "OLS Model 4: QQ-plot")
par(mfrow = c(1, 1))        # reset layout


#----Methodology 3.4----
#---3.4.1
# Direct post-stratification weighted mean calculation
library(dplyr)
library(forcats)

# Helper: recoding stratification variables to match the Eurostat reference grid
recode_strata <- function(df) {
  df %>%
    mutate(
      region         = factor(region,      1:4, c("North","South","West","East")),
      degurba        = factor(degurba,     1:3, c("Urban","Towns & Suburbs","Rural")),
      isced_hh       = factor(isced_hh,    1:3, c("Low","Medium","High")),
      sex            = factor(sex,         c("M","F")),
      samplingseason = factor(samplingseason, 1:4,
                              c("Spring","Summer","Autumn","Winter")))}
# Helper: computing per-observation weights from the reference grid
compute_weights <- function(df, ref_grid) {
  df %>%
    left_join(ref_grid, by = c("region","sex","samplingseason","degurba","isced_hh")) %>%
```

```
    group_by(region, sex, samplingseason, degurba, isced_hh) %>%
    mutate(obs_weight_raw = overall_weight / n()) %>%
    ungroup() %>%
    mutate(obs_weight = obs_weight_raw / sum(obs_weight_raw, na.rm = TRUE))}

# Preparing datasets and missingness check
# (a) restricting to age 6{12 & EU-27
full_data <- data_HBMC %>%
  filter(between(ageyears, 6, 12), region %in% 1:4)

# (b) dropping missing biomarker
missing_bio <- sum(is.na(full_data$mbzp_impcrtlog))
message(sprintf("Excluded %d records with missing MBzP", missing_bio))
data_clean <- full_data %>% filter(!is.na(mbzp_impcrtlog))

# (c) recoding factors & dropping stratification-missing
data_poststrat <- data_clean %>%
  recode_strata() %>%
  drop_na(region, sex, samplingseason, degurba, isced_hh)

message(sprintf("Post-stratification sample size: %d", nrow(data_poststrat)))

# Computing direct post-stratification weighted mean
weights_complete <- compute_weights(data_poststrat, ref_grid_base)

# check: distribution of observation weights
summary(weights_complete$obs_weight)

weighted_mean_complete <- sum(
  weights_complete$mbzp_impcrtlog * weights_complete$obs_weight,
  na.rm = TRUE)

#"Distribution of Post-stratification Weights"
# Computing percentiles on the original scale, then log-transform
pcts_d <- quantile(weights_complete$obs_weight,
                   probs = c(0.01, 0.05, 0.95, 0.99))

# plot
ggplot(weights_complete, aes(x = log10(obs_weight))) +
  geom_histogram(binwidth = 0.2,
                 fill   = "grey70",
                 colour = "black") +
  geom_vline(xintercept = log10(pcts_d),
             linetype   = "dashed",
             colour     = "red") +
  scale_x_continuous(
    breaks = seq(
      floor(min(log10(weights_complete$obs_weight))),
      ceiling(max(log10(weights_complete$obs_weight))),
      by = 1
    ),
    labels = function(x) sprintf("1e%+.0f", x)
  ) +
  labs(
```

```
    x     = expression(Log[10]~"Post-stratification weight"),
    y     = "Count",
    title = NULL) +
  theme_classic(base_size = 13)

# UNWEIGHTED geometric mean
# a. Unweighted on the full biomarker-complete sample (n = 2 784)
log_mean_full <- mean(data_clean$mbzp_impcrtlog, na.rm = TRUE)
geo_mean_full <- exp(log_mean_full)
cat(sprintf(
  "Full unweighted (n = %d): log-mean = %.4f → GM = %.2f µg/g crt\n",
  nrow(data_clean), log_mean_full, geo_mean_full))
# b. Unweighted on the post-stratification sample (n = 2 722)
log_mean_ps   <- mean(data_poststrat$mbzp_impcrtlog, na.rm = TRUE)
geo_mean_ps   <- exp(log_mean_ps)
cat(sprintf(
  "Restricted unweighted (n = %d): log-mean = %.4f → GM = %.2f µg/g crt\n\n",
  nrow(data_poststrat), log_mean_ps, geo_mean_ps))

# sensitivity analysis: imputing missing ISCED → "Medium"
weights_imp <- data_clean %>%
  recode_strata() %>%
  mutate(isced_hh = fct_na_value_to_level(isced_hh, "Medium")) %>%
  drop_na(degurba) %>%
  compute_weights(ref_grid_base)

weighted_mean_imp <- sum(
  weights_imp$mbzp_impcrtlog * weights_imp$obs_weight,
  na.rm = TRUE)

delta_pct <- 100 * (weighted_mean_imp - weighted_mean_complete) /
  weighted_mean_complete

# Report
geo_mean_complete <- exp(weighted_mean_complete)
geo_mean_imp      <- exp(weighted_mean_imp)

cat(sprintf(
  "Weighted geometric mean (complete) = %.2f µg g⁻¹  (log = %.4f)\n\
Weighted geometric mean (imputed)  = %.2f µg g⁻¹  (log = %.4f; Δ = %.2f%%)\n",
  geo_mean_complete,  weighted_mean_complete,
  geo_mean_imp,       weighted_mean_imp,      delta_pct))

#---3.4.2
# Survey-design estimation | log-scale mean, SE & GM + CI

library(survey)
# taking the post-stratified data + weights already made:
# data_poststrat (n=2722, with mbzp_impcrtlog) and weights_complete
weights_ps <- weights_complete
data_ps    <- data_poststrat

# building the survey design(s)
# independent (each child its own PSU)
```

```
des_ind <- svydesign(
  ids     = ~1,
  weights = ~obs_weight,
  data    = weights_ps,
  nest    = TRUE)

# helper: extracting log-scale mean & SE
log_est <- function(des) {
  m <- svymean(~mbzp_impcrtlog, des, na.rm = TRUE)
  c(mu = as.numeric(coef(m)), se = as.numeric(SE(m)))}

# helper: extracting geometric mean + 95% CI
gm_ci <- function(des) {
  L <- svymean(~mbzp_impcrtlog, des, na.rm = TRUE)
  mu <- as.numeric(coef(L))
  se <- as.numeric(SE(L))
  gm <- exp(mu)
  ci <- exp(mu + c(-1.96, +1.96) * se)
  c(GM = gm, L = ci[1], U = ci[2])}

# a. independent design
ind_log <- log_est(des_ind)
ind_gm  <- gm_ci(des_ind)

cat("=== Independent (no clustering) ===\n")
cat(sprintf("Log-scale mean  = %.4f (SE = %.4f)\n",
            ind_log["mu"], ind_log["se"]))
cat(sprintf("Geometric mean = %.2f µg/g crt  (95%% CI: %.2f{%.2f)\n\n",
            ind_gm["GM"], ind_gm["L"], ind_gm["U"]))

# b. clustered design (cohort as PSU), if you have a 'cohort' column
if ("cohort" %in% names(weights_ps)) {
  weights_ps$cohort <- factor(weights_ps$cohort)
  des_clu <- svydesign(
    ids     = ~cohort,
    weights = ~obs_weight,
    data    = weights_ps,
    nest    = TRUE)
  clu_log <- log_est(des_clu)
  clu_gm  <- gm_ci(des_clu)

  cat("=== Clustered (by cohort) ===\n")
  cat(sprintf("Log-scale mean  = %.4f (SE = %.4f)\n",
              clu_log["mu"], clu_log["se"]))
  cat(sprintf("Geometric mean = %.2f µg/g crt  (95%% CI: %.2f{%.2f)\n\n",
              clu_gm["GM"], clu_gm["L"], clu_gm["U"]))}

# c. trimming sensitivity
trim_report <- function(pct) {
  cap <- quantile(weights_ps$obs_weight, pct)
  df  <- weights_ps
  df$obs_weight <- pmin(df$obs_weight, cap)
  df$obs_weight <- df$obs_weight / sum(df$obs_weight)
  des_trim <- svydesign(ids=~1, weights=~obs_weight, data=df, nest=TRUE)
```

```
  t_gm <- gm_ci(des_trim)
  delta <- 100*(t_gm["GM"] - ind_gm["GM"])/ind_gm["GM"]
  cat(sprintf("Trim %2.0f%% (cap=%.4f): GM = %.2f (95%% CI %.2f{%.2f), Δ = %.2f%%\n",
              pct*100, cap, t_gm["GM"], t_gm["L"], t_gm["U"], delta))}

cat("=== Trimming sensitivity ===\n")
trim_report(0.99)
trim_report(0.95)

#=======additional paragraph on sensitivity calibration for regionxage=======
##  SENSITIVITY: add a REGION × AGE margin to the 5-way post-strat weights  ##
## clustered \five-way" post-stratified design (baseline)
dat_ageSens <- data_ps %>%                # ← rows n = 2 722, already recoded
  mutate(
    w_post = weights_ps$obs_weight,       # probability weights (Σ = 1)
    cohort = factor(cohort))               # 11 PSUs

N_ageSens        <- nrow(dat_ageSens)     # sample size on the frequency scale
dat_ageSens$w_pop <- dat_ageSens$w_post * N_ageSens   # Σw_pop = n

des_post_clu_ageSens <- svydesign(
  ids     = ~cohort,
  weights = ~w_pop,                       # five-way post-strat FREQ weights
  data    = dat_ageSens,
  nest    = TRUE)

## Target table: REGION × AGE  (uniform 6{12 within each region)
# only uses strata that exist in the sample
# ensures ΣFreq = n  and  ΣFreq_region = region's sample size
# A. distinct region{age strata observed in the data
obs_cells <- dat_ageSens %>%
  distinct(region, ageyears)
# B. frequency-scale sample size per region
region_totals <- dat_ageSens %>%
  count(region, name = "region_N")        # Σregion_N = N_ageSens
# C. distribute each region's total equally across its observed age strata
pop_ageReg_ageSens <- obs_cells %>%
  left_join(region_totals, by = "region") %>%
  group_by(region) %>%
  mutate(Freq = region_N / n()) %>%       # n() = # age-years present
  ungroup() %>%
  select(region, ageyears, Freq)

## check
stopifnot(abs(sum(pop_ageReg_ageSens$Freq) - N_ageSens) < 1e-6)

## raking the design to the REGION × AGE margin (keeps cohort PSUs)
des_ageReg_ageSens <- rake(
  design             = des_post_clu_ageSens,
  sample.margins     = list(~region + ageyears),
  population.margins = list(pop_ageReg_ageSens),
  control            = list(maxit = 200, epsilon = 1e-6))

## Age-standardised EU geometric mean  + 95 % CI
```

```
gm_obj <- svymean(~mbzp_impcrtlog, des_ageReg_ageSens)
mu_log <- coef(gm_obj)[1]
se_log <- SE(gm_obj)[1]

GM_age <- exp(mu_log)
CI_age <- exp(mu_log + qnorm(c(.025, .975)) * se_log)

## 4Weight-only DEFF and effective n
prob_w <- weights(des_ageReg_ageSens) / sum(weights(des_ageReg_ageSens))
DEFF   <- N_ageSens * sum(prob_w^2)
n_eff  <- round(N_ageSens / DEFF)

## Comparison with the main five-way post-strat × cohort estimate
GM_main <- exp(coef(svymean(~mbzp_impcrtlog, des_post_clu_ageSens))[1])
delta   <- 100 * (GM_age - GM_main) / GM_main

## Concise summary
cat("\n===== AGE-CALIBRATED EU MBzP (REGION × AGE, 11 COHORTS) =====\n")
cat(sprintf("Geometric mean        = %.2f µg g⁻¹ crt  (95 %% CI %.2f { %.2f)\n",
            GM_age, CI_age[1], CI_age[2]))
cat(sprintf("Weight-only DEFF      = %.2f  →  n_eff  %d\n", DEFF, n_eff))
cat(sprintf("Change vs main (%.2f) = %.1f %%\n", GM_main, delta))

## AGE-CALIBRATED (REGION × AGE) { UNCLUSTERED SENSITIVITY
## Five-way post-strat design *without* PSUs
des_post_ind <- svydesign(
  ids     = ~1,                          # <-- independence
  weights = ~w_pop,                      # same frequency weights Σ = n
  data    = dat_ageSens,                 # n = 2 722
  nest    = TRUE)

## REGION × AGE target (same pop_age, Reg_ageSens as before)
## (code identical)

## Calibration
des_ageReg_ind <- rake(
  design             = des_post_ind,
  sample.margins     = list(~region + ageyears),
  population.margins = list(pop_ageReg_ageSens),
  control            = list(maxit = 200, epsilon = 1e-6))

## GM & CI
gm   <- svymean(~mbzp_impcrtlog, des_ageReg_ind)
gm_u <- exp( coef(gm)[1] )
ci_u <- exp( coef(gm)[1] + qnorm(c(.025,.975))*SE(gm)[1] )

## weight-only DEFF
p_w  <- weights(des_ageReg_ind)/sum(weights(des_ageReg_ind))
deff <- N_ageSens*sum(p_w^2);   n_eff <- round(N_ageSens/deff)

cat(sprintf(
  "\nUNclustered age-raked GM = %.2f µg/g crt  (95%% CI %.2f { %.2f)\n",
  gm_u, ci_u[1], ci_u[2]))
cat(sprintf("Weight-only DEFF = %.2f  → n_eff  %d\n", deff, n_eff))
```

```
#---3.4.3
#RAKING (marginal calibration to Eurostat proportions)
## Base (unclustered, unit-weight) design
rake_design_base <- svydesign(ids = ~1, weights = ~1, data = data_ps)

## Helper to guarantee factor-level consistency
check_levels <- function(var, pop_tbl) {
  if (!identical(levels(data_ps[[var]]), levels(pop_tbl[[var]])))
    stop(sprintf("Level mismatch in '%s'", var), call. = FALSE)}

## Eurostat targets, re-scaled to N
N <- nrow(data_ps)
rake_pop_region  <- data.frame(region       = external_region$region,
                               Freq         = external_region$region_prop  * N)
rake_pop_sex     <- data.frame(sex          = external_sex$sex,
                               Freq         = external_sex$sex_prop       * N)
rake_pop_season  <- data.frame(samplingseason = external_season$samplingseason,
                               Freq           = external_season$season_prop * N)
rake_pop_degurba <- data.frame(degurba       = external_degurba$DEGURBA,
                               Freq          = external_degurba$degurba_prop * N)
rake_pop_isced   <- data.frame(isced_hh      = external_isced$isced_hh,
                               Freq          = external_isced$isced_prop    * N)

## Check that every factor in the sample has the same levels
## as the corresponding Eurostat table
lapply(list(region         = rake_pop_region,
            sex            = rake_pop_sex,
            samplingseason = rake_pop_season,
            degurba        = rake_pop_degurba,
            isced_hh       = rake_pop_isced),
       \(tbl) check_levels(names(tbl)[1], tbl))

## raking (iterative proportional fitting)
rake_design <- rake(
  design             = rake_design_base,
  sample.margins     = list(~region, ~sex, ~samplingseason, ~degurba, ~isced_hh),
  population.margins = list(rake_pop_region, rake_pop_sex, rake_pop_season,
                            rake_pop_degurba, rake_pop_isced))

# Weight diagnostics (probability-scale DEFF, etc.)
freq_w <- weights(rake_design)          # Σ = N

data_ps$obs_weight <- freq_w / sum(freq_w)   # probability weights Σ = 1

prob_w <- data_ps$obs_weight
cat("Probability-weight range :",
    formatC(range(prob_w), format = "e", digits = 2), "\n")

deff_w <- N * sum(prob_w^2)                # weights-only DEFF
n_eff  <- N / deff_w
cat("Weights-only DEFF        =", round(deff_w, 2),
    "(n_eff ", round(n_eff), ")\n\n")
```

```
## Point estimate on log scale and back-transform
rake_est <- svymean(~mbzp_impcrtlog, rake_design)
mu_log   <- coef(rake_est)[1]
se_log   <- SE(rake_est)[1]
gm       <- exp(mu_log)
gm_ci    <- exp(mu_log + qnorm(c(.025, .975)) * se_log)

cat(sprintf("Log scale: μ = %.4f (SE = %.4f)\n", mu_log, se_log))
cat(sprintf("Raked GM  = %.2f μg/g crt  (95%% CI: %.2f{%.2f)\n",
            gm, gm_ci[1], gm_ci[2]))

## Cluster-robust CI (cohort as PSU)
if ("cohort" %in% names(data_ps)) {

data_ps$rake_wt    <- freq_w            # frequency weights Σ = N
# probability weights already in data_ps$obs_weight
rake_design_cl <- svydesign(ids = ~cohort, weights = ~rake_wt,
                            data = data_ps, nest = TRUE)

cl_est <- svymean(~mbzp_impcrtlog, rake_design_cl)
cl_ci  <- exp(coef(cl_est)[1] + c(-1.96, 1.96) * SE(cl_est)[1])

cat(sprintf("Clustered GM CI       = %.2f { %.2f μg/g crt\n",
              cl_ci[1], cl_ci[2]))}

## Histogram of raked weights (probability scale)
pct <- quantile(prob_w, probs = c(.01, .05, .95, .99))
ggplot(data.frame(w = prob_w), aes(x = log10(w))) +
  geom_histogram(binwidth = 0.2, fill = "grey70", colour = "black") +
  geom_vline(xintercept = log10(pct), linetype = "dashed", colour = "red") +
  scale_x_continuous(breaks = seq(floor(min(log10(prob_w))),
                                  ceiling(max(log10(prob_w))), 1),
                     labels = function(x) sprintf("1e%+.0f", x)) +
  labs(x = expression(Log[10]~"Raked probability weight"),
       y = "Count") +
  theme_classic(base_size = 13)

#Section 4.1
# ===EDA===
library(dplyr)
library(ggplot2)

#BOXPLOTS
# EDA dataset: dropping only missing outcome, keep ages 6{12 & EU-27 regions
eda_all <- data_HBMC %>%
  filter(
    between(ageyears, 6, 12),
    region %in% 1:4,
    !is.na(mbzp_impcrtlog)
  ) %>%
  recode_strata()   # turns region, sex, samplingseason, degurba, isced_hh into factors

# Boxplot by Region
```

```
ggplot(eda_all %>% filter(!is.na(region)),
       aes(x = region, y = mbzp_impcrtlog)) +
  geom_boxplot(outlier.shape = 21, outlier.fill = "white") +
  labs(
    x     = "Region",
    y     = expression(log~MBzP~(µg/g~creatinine)),
    title = NULL
  ) +
  theme_minimal()

data_ps %>% group_by(region) %>% summarise(
  med = median(mbzp_impcrtlog),
  Q1  = quantile(mbzp_impcrtlog, .25),
  Q3  = quantile(mbzp_impcrtlog, .75),
  min = min(mbzp_impcrtlog),
  max = max(mbzp_impcrtlog))

# similar for DEGURBA, Household ISCED, sampling season and sex boxplots

#sampling years and cohorts
library(forcats)
library(viridis)   # for scale_fill_viridis_c()
# ordering countries by their first sampling year
country_order <- dataHBMC_plots %>%
  group_by(country) %>%
  summarise(first_year = min(samplingyear), .groups = "drop") %>%
  arrange(first_year) %>%
  pull(country)

# heat-map
dataHBMC_plots %>%
  count(country, samplingyear) %>%
  mutate(country = factor(country, levels = country_order)) %>%
  ggplot(aes(x = samplingyear, y = country, fill = n)) +
  geom_tile(colour = "white") +
  scale_x_continuous(breaks = 2014:2021, labels = 2014:2021,
                     expand = c(0, 0)) +
  scale_fill_viridis_c(name = "Participants") +
  labs(title = NULL,
       x = "Year", y = "Country") +
  theme_minimal(base_size = 14) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    panel.grid  = element_blank())

#COHORT violion plot
# Violin + boxplot overlay
ggplot(eda_all %>% filter(!is.na(cohort)),
       aes(x = factor(cohort),
           y = mbzp_impcrtlog,
           fill = factor(cohort))) +
  geom_violin(trim = TRUE, alpha = 0.6, color = NA) +
  geom_boxplot(width = 0.1, outlier.shape = NA,
               fill = "white", color = "black") +
```

```
  scale_fill_brewer(palette = "Set3") +
  labs(
    x     = "Cohort",
    y     = expression(log~MBzP~(µg/g~creatinine)),
    title = NULL
  ) +
  theme_classic() +
  theme(
    axis.text.x     = element_text(angle = 45, hjust = 1),
    legend.position = "none")

eda_all %>%
  filter(!is.na(cohort)) %>%
  group_by(cohort) %>%
  summarise(
    n    = n(),
    med  = median(mbzp_impcrtlog),
    Q1   = quantile(mbzp_impcrtlog, .25),
    Q3   = quantile(mbzp_impcrtlog, .75),
    min  = min(mbzp_impcrtlog),
    max  = max(mbzp_impcrtlog))

#As for HBM4EU comparisons vs Eurostat data
#REGION distribution of the sample
library(scales)

# sample counts by Region
sample_region <- eda_all %>%
  count(region) %>%
  mutate(count_m = n / 1e6)
# pastel palette
cols <- c(
  West  = "#DECBE4",
  North = "#CBD5E8",
  East  = "#B3E2CD",
  South = "#FDBF6F")

# plotting raw counts
ggplot(sample_region, aes(region, n, fill = region)) +
  geom_col(width = 0.7) +
  scale_y_continuous(
    labels = comma,
    expand = expansion(mult = c(0, 0.05))
  ) +
  scale_fill_manual(values = cols, guide = "none") +
  labs(
    x = NULL,
    y = "HBM4EU: number of children (6-12 yrs)",
    title = NULL
  ) +
  theme_minimal(base_size = 13) +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
```

```
    panel.background = element_blank())

sample_region <- eda_all %>%
  count(region) %>%
  mutate(
    pct = round(n / sum(n) * 100, 1))

#AGE distribution within HBM4EU
# Preparing sample age{region distributions
sample_age_region <- eda_all %>%
  group_by(region, ageyears) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(region) %>%
  mutate(prop = count / sum(count)) %>%
  ungroup()

# Plot
ggplot(sample_age_region, aes(x = factor(ageyears), y = prop, fill = region)) +
  geom_col(width = 0.8) +
  facet_wrap(~ region, nrow = 2) +
  scale_fill_manual(values = cols, guide = "none") +
  scale_y_continuous(
    labels = percent_format(1),
    expand = expansion(mult = c(0, 0.05))
  ) +
  labs(
    x = "Age (years)",
    y = "% within region",
    title = NULL
  ) +
  theme_minimal(base_size = 13) +
  theme(
    panel.grid.major  = element_blank(),
    panel.grid.minor  = element_blank(),
    strip.background  = element_rect(fill = "white", colour = NA),
    strip.text        = element_text(face = "bold"),
    axis.text.x       = element_text(size = 10))

sample_age_region <- eda_all %>%
  group_by(region, ageyears) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(region) %>%
  mutate(pct = round(n / sum(n) * 100, 1)) %>%
  ungroup()

#DEGURBA distribution within HBM4EU
# Aggregating sample by region × DEGURBA
sample_region_deg <- eda_all %>%
  filter(!is.na(degurba)) %>%
  group_by(region, degurba) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(region) %>%
  mutate(
    pct         = n / sum(n),        # share within region
```

```
    n_thousands = n / 1000           # count in thousands
  ) %>%
  ungroup()

# plotting raw counts (in thousands), same with Eurostat palette
euro_palette <- c(
  "Urban"            = "lightblue",
  "Towns & Suburbs"  = "lightgreen",
  "Rural"            = "lightpink")

# Plotting percentages with the same Eurostat palette
ggplot(sample_region_deg, aes(x = region, y = pct, fill = degurba)) +
  geom_col(position = "dodge", width = 0.7) +
  scale_y_continuous(
    labels = percent_format(1),
    expand = expansion(mult = c(0, 0.05))
  ) +
  scale_fill_manual(values = euro_palette) +
  labs(
    x     = NULL,
    y     = "Share (%) within region",
    title = NULL
  ) +
  theme_minimal(base_size = 13) +
  theme(
    panel.grid   = element_blank(),
    legend.title = element_blank(),
    axis.text.x  = element_text(angle = 45, hjust = 1))

degurba_summary <- sample_region_deg %>%
  mutate(
    pct = round(pct * 100, 1)        # convert to percent with 1 decimal
  ) %>%
  select(region, degurba, n, pct)

#ISCED distribution within HBM4EU
# Summary sample by region × ISCED
sample_region_edu <- eda_all %>%
  filter(!is.na(isced_hh)) %>%
  count(region, isced_hh) %>%
  group_by(region) %>%
  mutate(pct = n / sum(n)) %>%
  ungroup()
print(sample_region_edu)

sample_region_edu <- eda_all %>%
  filter(!is.na(isced_hh)) %>%
  count(region, isced_hh) %>%
  group_by(region) %>%
  mutate(pct = n / sum(n)) %>%
  ungroup() %>%
  # re-ordering region and education levels
  mutate(
    region   = factor(region,   levels = c("North","South","West","East")),
```

```
    isced_hh = factor(isced_hh, levels = c("High","Medium","Low")))

ggplot(sample_region_edu, aes(region, pct, fill = isced_hh)) +
  geom_col(position = "dodge", width = 0.7) +
  scale_x_discrete(limits = c("North","South","West","East")) +   # ensures x-axis order
  scale_y_continuous(labels = scales::percent_format(1), expand = expansion(mult = c(0,0.05))) +
  scale_fill_manual(
    values = c("High"="#ccebc5","Medium"="#b3cde3","Low"="#fbb4ae"),
    name = "Education level"
  ) +
  labs(
    x = NULL,
    y = "Share within region",
    title = NULL
  ) +
  theme_minimal(base_size = 13) +
  theme(
    panel.grid  = element_blank(),
    axis.text.x = element_text(angle = 45, hjust = 1))

#age-outcome scatter + (LOESS plot)
# ageyears numeric
eda_all <- eda_all %>% mutate(ageyears = as.numeric(as.character(ageyears)))

ggplot(eda_all, aes(x = ageyears, y = mbzp_impcrtlog)) +
  geom_point(alpha = 0.3, size = 1) +
  geom_smooth(method = "loess", span = 0.75, se = TRUE) +
  labs(
    x = "Age (years)",
    y = expression(log~MBzP~(µg/g~creatinine))
  ) +
  theme_classic(base_size = 13)

#additional data for age-outcome HBM4EU
#Summary stats by age
age_summary <- eda_all %>%
  group_by(ageyears) %>%
  summarise(
    n        = n(),
    mean     = mean(mbzp_impcrtlog, na.rm = TRUE),
    median   = median(mbzp_impcrtlog, na.rm = TRUE),
    sd       = sd(mbzp_impcrtlog, na.rm = TRUE),
    Q1       = quantile(mbzp_impcrtlog, 0.25, na.rm = TRUE),
    Q3       = quantile(mbzp_impcrtlog, 0.75, na.rm = TRUE))

# Linear model (log-MBzP ~ age)
lm_age <- lm(mbzp_impcrtlog ~ ageyears, data = eda_all)
summary(lm_age)
# Quadratic model to test nonlinearity
lm_age2 <- lm(mbzp_impcrtlog ~ ageyears + I(ageyears^2), data = eda_all)
anova(lm_age, lm_age2)

#Reference-Grid Cell Weights on a log-scale
#Numeric summary of grid weights
```

```
summ_w <- ref_grid_base %>%
  summarise(
    min    = min(overall_weight),
    p1     = quantile(overall_weight, 0.01),
    p5     = quantile(overall_weight, 0.05),
    median = median(overall_weight),
    mean   = mean(overall_weight),
    p95    = quantile(overall_weight, 0.95),
    p99    = quantile(overall_weight, 0.99),
    max    = max(overall_weight),
    N_heavy= sum(overall_weight > 0.01))
print(round(summ_w, 5))

# percentiles
qs <- ref_grid_base %>%
  pull(overall_weight) %>%
  quantile(c(0.01, 0.05, 0.95, 0.99))

#Plot with dashed lines at percentiles
ggplot(ref_grid_base, aes(x = log10(overall_weight))) +
  geom_histogram(binwidth = 0.1, fill = "grey70", color = "black") +
  geom_vline(
    xintercept = log10(qs),
    linetype   = "dashed",
    color      = "red"
  ) +
  scale_x_continuous(
    breaks = seq(
      floor(min(log10(ref_grid_base$overall_weight))),
      ceiling(max(log10(ref_grid_base$overall_weight))),
      by = 1),
    labels = function(x) sprintf("1e%+d", x)) +
  labs(
    x = expression(Log[10]~Cell~Weight),
    y = "Count") +
  theme_classic()
```