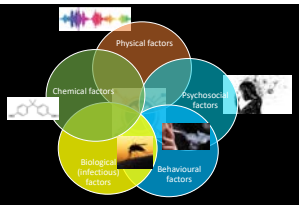


The Exposome: Promises and Challenges of a New Concept

L'exposome : Promesses et défis d'un nouveau concept

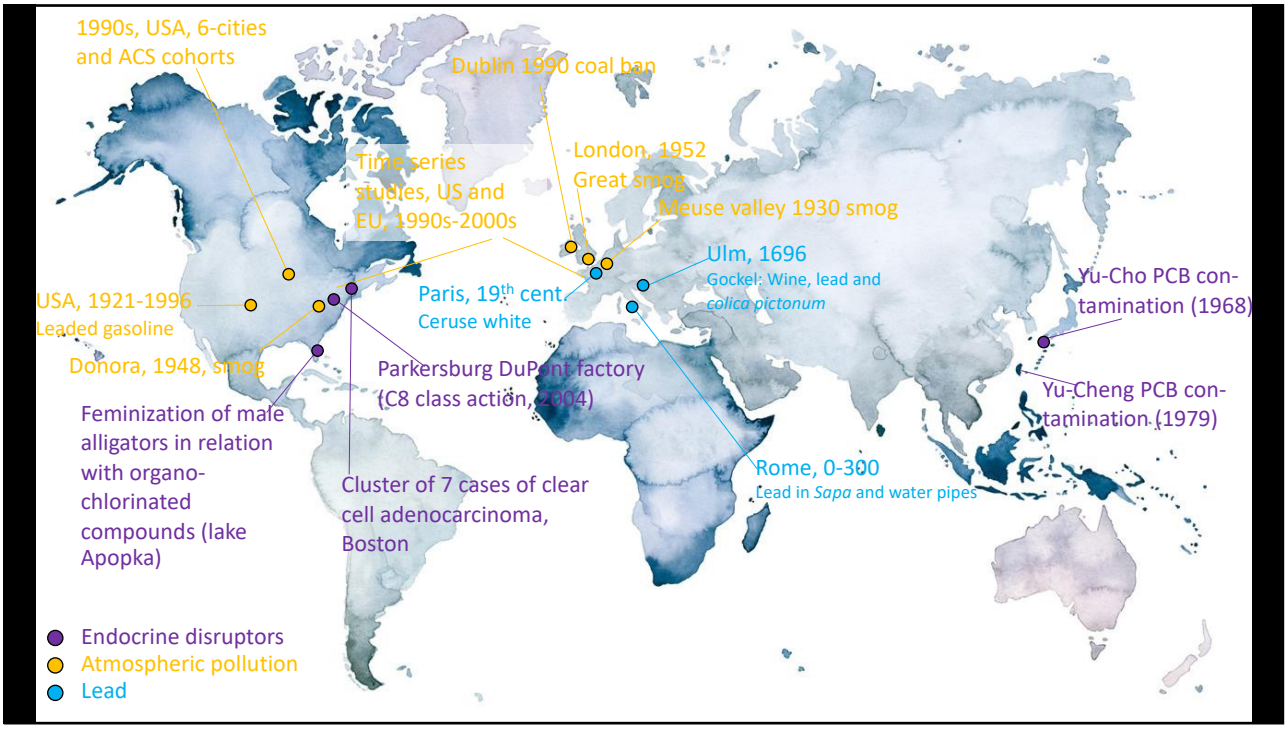
Rémy Slama
Collège de France & Inserm

The relations between human health and the environment in the Anthropocene

Lecture #7 – 25 May 2022

COLLÈGE DE FRANCE 1530

1



2

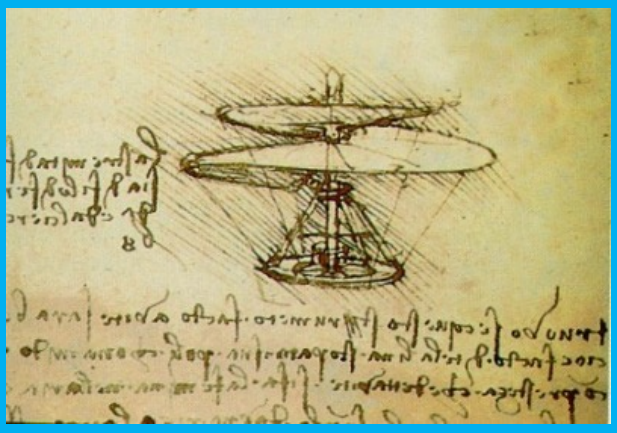
Lecture overview

A. Motivation, aims & challenges
B. Layer 1: exposome descriptive studies (biomonitoring / environmental justice challenge)
C. 2-layer problem: exposome-health studies
D. Multi-layer problem: Cross-omics analyses
E. Perspectives: the first 20 years of exposome research

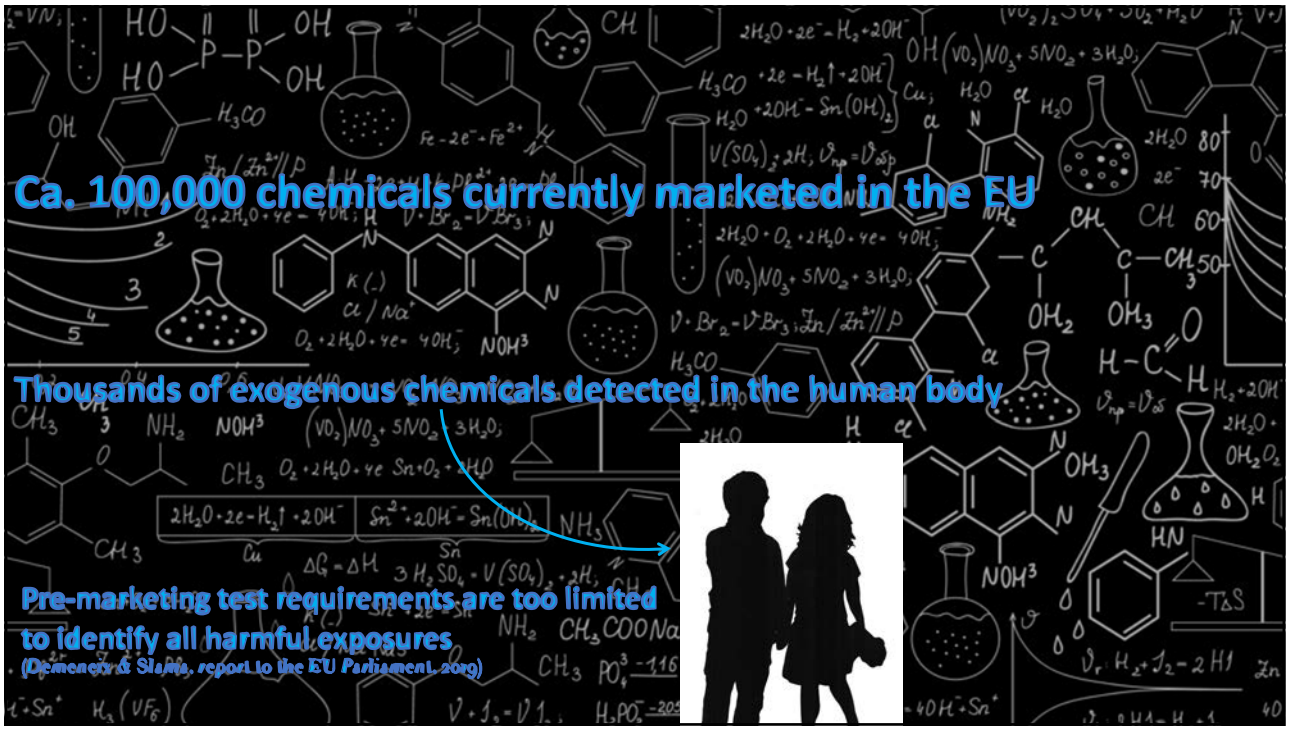
The French biomonitoring survey *(seminar of Dr. Clémence Fillol)*

3

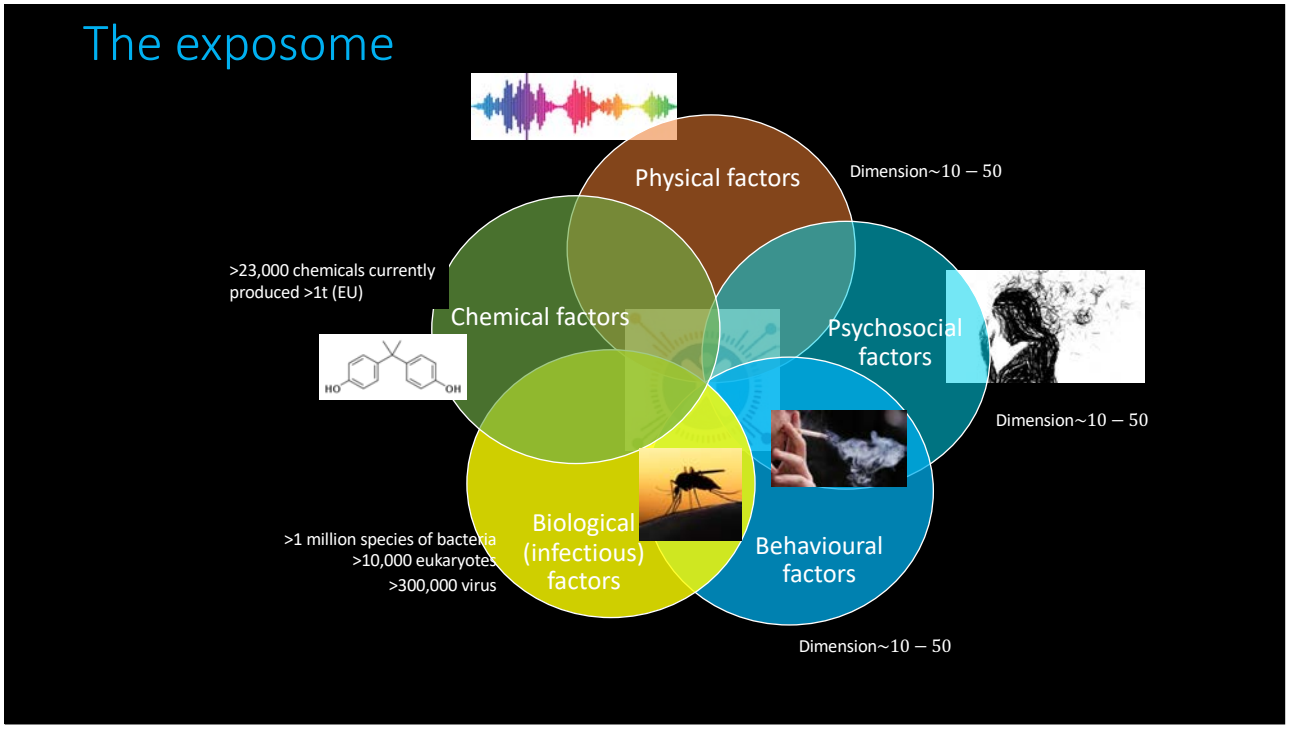
A. Motivation, aims & challenges of exposome research



4

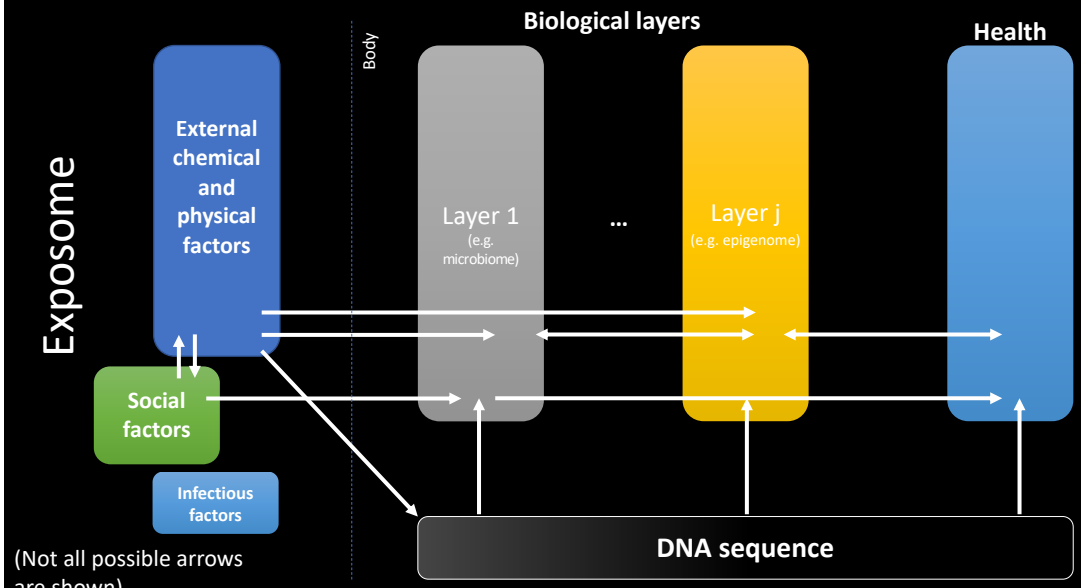

Ca. 100,000 chemicals currently marketed in the EU
Thousands of exogenous chemicals detected in the human body
Pre-marketing test requirements are too limited to identify all harmful exposures
(Demeneix & Slama, report to the EU Parliament, 2019)

5


The exposome
Physical factors
Dimension~10 – 50
>23,000 chemicals currently produced >1t (EU)
Chemical factors
Psychosocial factors
Dimension~10 – 50
>1 million species of bacteria
>10,000 eukaryotes
>300,000 virus
Biological (infectious) factors
Behavioural factors
Dimension~10 – 50

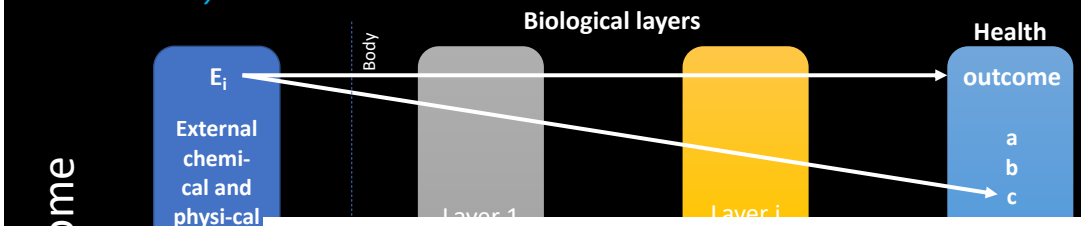
6

From the environment to human health: *an analytical model*



7

From the environment to human health: *an analytical model*



Some possible questions

1) What happens to health outcome c if E_i varies? (e.g., : does smoking cause lung cancer?) [**classical, single exposure, aetiology**]
2) What are all the external risk factors that influence c? [exposome-health study/**exposome-wide study**]
3) What happens to health if E_i varies? (e.g., : what are the sanitary consequences of smoking?) [**outcome-wide aetiology**]
4) What intermediary variables may explain an effect of E_i on c? [**Mechanistic research**/mediation analysis]
5) Which risk factor has the largest *impact* (i.e. attributable number of disease cases) on c today? [**Environmental burden of disease**] On health overall?
6) Are specific sociodemographic subgroups disproportionately exposed to harmful exposures? [**Environmental justice problem**]

8

Question 1: What happens to health outcome c if E_i varies? [classical aetiology]

The answer of « classical » risk factor epidemiology *(assuming a randomized controlled experiment is not feasible, e.g. for ethical reasons):*

- Recruit a cohort of "healthy" subjects
- Assess E (at inclusion and possibly repeatedly)
- Assess sociodemographic and behavioural factors also potentially influencing c
- Wait for outcome c to occur in a large enough number of subjects of the cohort
- Quantify the statistical relation between E and c adjusting for the "potential confounders"
- Repeat the study a certain number of times
- Perform a meta-analysis of all published studies and report to decision makers (if possible incorporating external mechanistic evidence and quantifying the overall level of evidence)

Doll & Hill's *Doctors' study*

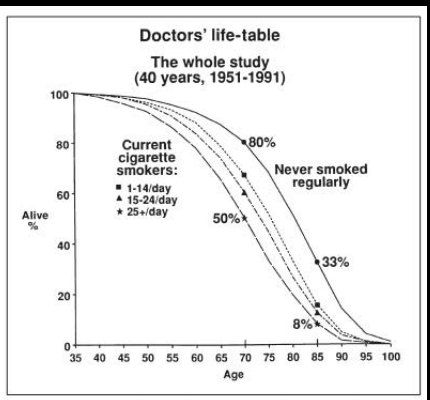


Figure 7 Survival of British doctors from age 35 years by smoking habits: lifelong nonsmokers —, cigarette smokers, smoking 1–14 cigarettes a day ·····, smoking 15–24 cigarettes a day – · –, 25 or more cigarettes a day – – –[18] (Reproduced with permission of the BMJ Publishing Group)

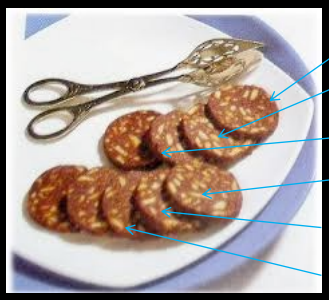
9

Problems with the approach of classical risk factors epidemiology

- E_i may be measured with error (**measurement error**)
- Confounders (and outcome) may be measured with error
- There may be confounding by other unmeasured exposures (**confounding by co-exposures**)
- Effect-measure modifications ("**interactions/synergy**") not easily studied (unless strong a priori hypothesis)
- There may be hidden **multiple testing** (e.g., because actually E_i was not an a priori choice)
- **Random fluctuations**
- **Rather *low throughput*** approach compared e.g. to toxicology (1 exposure every 1-2 years per PI in a given cohort?)
- The study is more likely to be published if it highlights a "significant" association, biasing the meta-analysis (**publication bias**)

10

Selling the salami by the slice…
(i.e., studying each exposure separately in the same cohort)



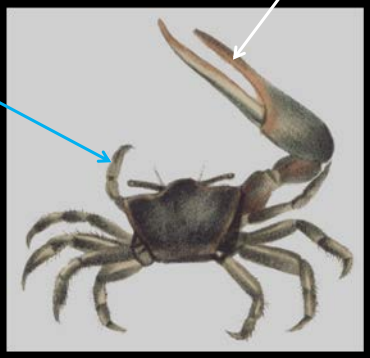
Risk of selective reporting, publication bias, chance finding through multiple-testing…

11

Characterizing the genetic and environmental influences on human health: unbalanced efforts

Effort in assessment of health and genetic factors

Effort in exposure assessment



Crab *Uca pugnax* (Wild, *Cancer Epid Biomarkers Prev*, 2005)

12

Deciphering the exposome: Motivation

There is a desperate need to develop methods with the same precision for an individual's environmental exposure as we have for the individual's genome. I would like to suggest that there is need for an ''exposome'' to match the ''genome.''

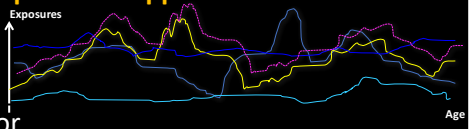
At its most complete, **the exposome encompasses life-course environmental exposures (including lifestyle factors), from the prenatal period onwards**.

(Wild CP, *Cancer Epid Biom Prev*, 2005)

14

Problems with the approach of classical risk factors epidemiology

Can (could) be taken care of in an "exposome" approach



- E_i may be measured with error (**measurement error**)
- Confounders (and outcome) may be measured with error
- There may be confounding by other unmeasured exposures (**confounding by co-exposures**)
- Effect-measure modifications ("**interactions/synergy**") not easily studied (unless strong a priori hypothesis)
- There may be hidden **multiple testing** (e.g., because actually E_i was not an a priori choice)
- **Random fluctuations**
- **Rather *low throughput*** approach compared e.g. to toxicology (1 exposure every 1-2 years per PI in a given cohort?)
- The study is more likely to be published if it highlights a "significant" association, biasing the meta-analysis (**publication bias**)

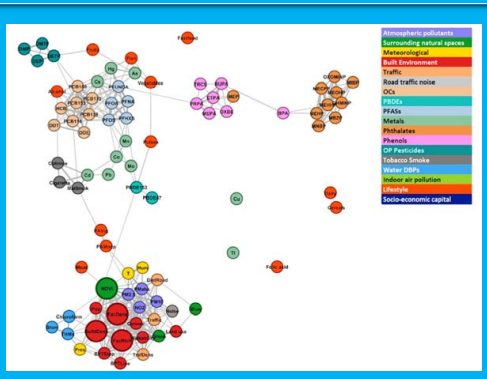
15

Challenges of exposome studies

1. Exposure assessment: width *vs.* accuracy trade-off (can we assess more compounds without increasing exposure misclassification?)
2. Associating the exposome and health: multiple-testing issues (can we consider more exposures without increasing too much bias and random error, while maintaining the ambition of deciphering causal relations?)
3. Cross-exposure (mixtures/synergy) and cross-omics analyses

18

B. Single-layer problem: describing the exposome



« There is a desperate need to develop methods **with the same precision for an individual's environmental exposure** as we have for the individual's genome. »

(Wild, *Canc Epid Biomark Prev*, 2005)

19

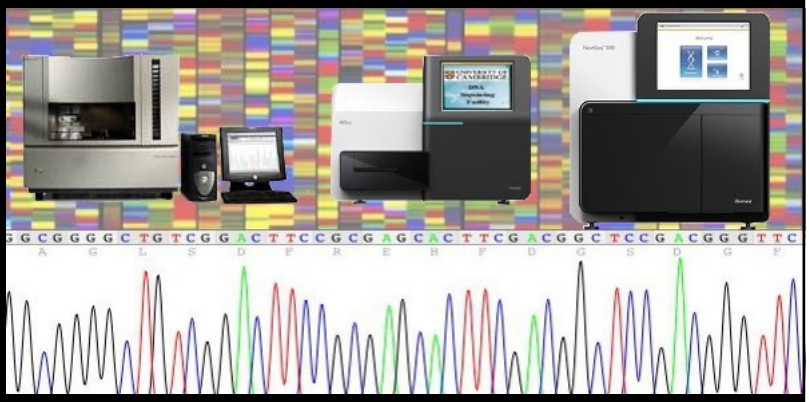
Metrologic issue

Increasing the number of (exposure) factors considered should not be done at the cost of a decrease in the quality of their assessment.
(cf. *curse of dimensionality* data science concept)

More exposures, better characterized

20

Burden for the typical participant to a *genome* study



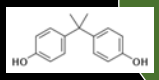
Simple alphabet composed of 4 letters (nucleotides): A, T, G, C
PCR techniques, DNA sequencing facilities
Cost of assessment of genetic polymorphisms: ~100€ per sample

21

The exposome

Physical factors

Chemical factors



Psychosocial factors

Biological (infectious) factors

Behavioural factors

22

Commentary

(wileyonlinelibrary.com) DOI: 10.1002/sim.5499

Published online in W

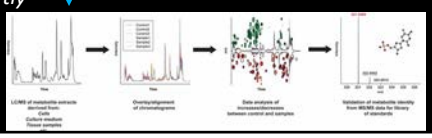
The biomarker revolution

Enrique F. Schisterman[a*†] and Paul S. Albert[b]

Targeted analyses (hundreds of chemicals), *very sensitive, annotated*

5 ml

Untargeted analysis (thousands of chemicals, *not always quantitative, partly annotated*)



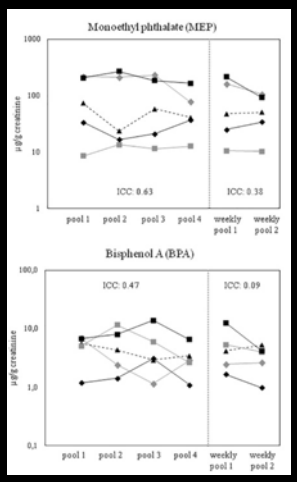
Compound		Maternal samples P25	P50	P75	Max	n analyzed	% quantifiable samples
PCB 118	ng/g lipid	1.57	2.64	4.82	39.0	829	79.1
PCB 138	ng/g lipid	5.54	9.10	16.1	132	1048	96.5
PCB 153	ng/g lipid	10.4	17.6	30.5	214	1048	99.6
PCB 170	ng/g lipid	1.84	3.69	7.06	84.5	826	99.5
PCB 180	ng/g lipid	5.78	10.4	18.6	201	1048	97.6
DDT	ng/g lipid	0.82	1.33	3.06	94.1	826	65.6
DDE	ng/g lipid	25.9	52.3	111	1903	1048	99.9
HCB	ng/g lipid	5.59	8.16	13.0	164	1048	99.1
PBDE 47	ng/g lipid	0.27	0.43	0.75	34.74	684	80.8
PBDE 153	ng/g lipid	0.03	0.45	0.66	198	648	72.9
PFOA	µg/L	1.38	2.30	3.34	31.6	1240	99.7
PFNA	µg/L	0.43	0.69	1.08	5.92	1240	97.9
PFUnDA	µg/L	0.10	0.19	0.29	2.80	1032	95.4
PFHxS	µg/L	0.31	0.55	0.91	21.0	1240	97.5
PFOS	µg/L	4.12	6.41	9.63	48.0	1240	100
As	µg/L	0.28	1.19	2.27	90.1	833	58.5
Cd	µg/L	0.15	0.22	0.33	27.9	833	99.6
Cs	µg/L	1.19	1.56	2.15	10.9	833	100
Hg	µg/L	0.98	1.90	3.45	43.5	1020	98.9
Pb	µg/L	7.14	9.66	13.20	187	833	100
Tl	µg/L					833	1.1
MEP	µg/g creat.	72.0	179	469	17,733	1080	99.0
MiBP	µg/g creat.	23.3	38.7	60.7	705	1088	99.9
MnBP	µg/g creat.	18.3	29.6	47.3	6445	1089	100
MBzP	µg/g creat.	3.63	7.33	15.3	775.1	1088	99.7
MEHP	µg/g creat.	4.42	8.73	15.3	417	1085	99.5
MEHHP	µg/g creat.	10.5	18.2	31.2	967	1089	100
MEOHP	µg/g creat.	8.29	14.1	23.7	783	1089	100
MECPP	µg/g creat.	22.4	33.6	52.3	1361	913	99.9
oh-MiNP	µg/g creat.	0.61	0.91	1.47	66.5	914	92.6
oxo-MiNP	µg/g creat.	0.62	1.03	1.75	75.1	914	95.7
MEPA	µg/g creat.	39.5	167	389	39,241	815	99.8
ETPA	µg/g creat.	1.14	6.26	26.72	6774	817	97.4
PRPA	µg/g creat.	8.87	44.2	134	12,463	1083	97.3
BUPA	µg/g creat.	0.36	3.37	14.4	371	1083	97.0
BPA	µg/g creat.	1.55	2.82	6.60	107	1084	99.4
OXBE	µg/g creat.	1.46	4.90	27.5	12,837	1085	99.3
TCS	µg/g creat.	1.50	6.28	79.9	1653	1085	98.5
DMP	µg/g creat.	4.13	8.37	16.4	321	1080	90.8
DMTP	µg/g creat.	2.05	4.96	12.4	220	1084	88.9
DMDTP	µg/g creat.	0.05	0.19	1.54	134	969	41.6
DEP	µg/g creat.	1.86	3.33	6.44	198	1082	97.8
DETP	µg/g creat.	0.12	0.58	2.56	44.3	1037	50
DEDTP	µg/g creat.					1084	1.7
Cotinine	µg/g creat.	1.49	5.25	23.7	15,410	1093	43.7

(Haug, *Env Int*, 2018)

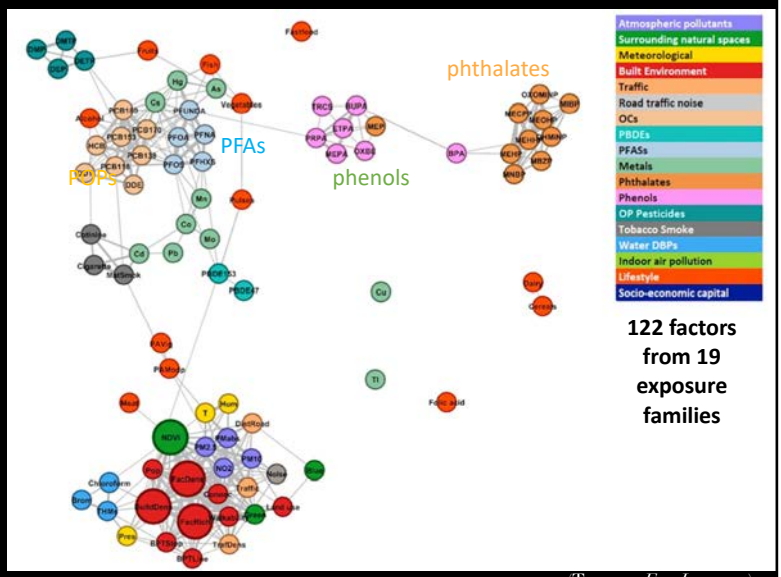
23

Characterization of ***correlations*** in the exposome

Within-subject (temporal variability) or *between exposures at a given time*



(Casas, *Env Int*, 2018)
See also (Vernet, *EHP*, 2018)

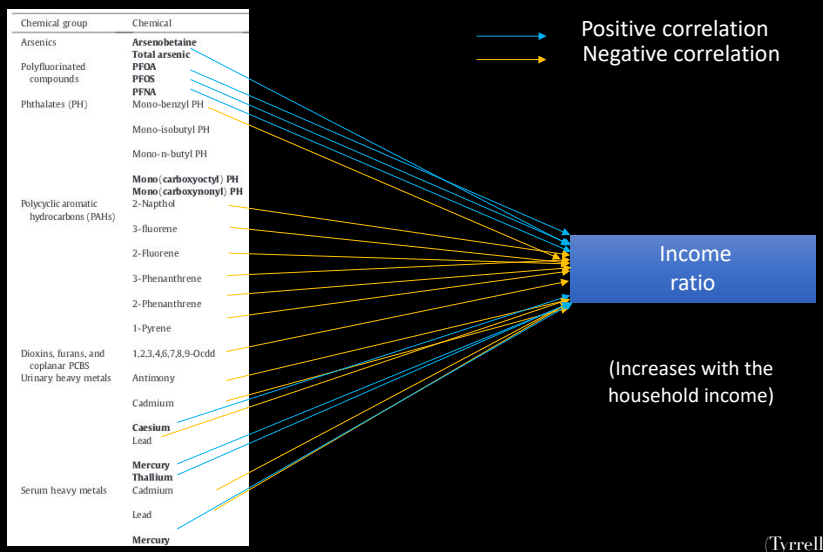


122 factors from 19 exposure families

(Tamayo, *Env Int*, 2019)

24

Exposome-wide social exposure contrasts: *Informing **environmental justice (1): USA (NHANES)***



(Tyrrell, *Env Int*, 2013)

25

Exposome-wide social exposure contrasts: *Informing **environmental justice (2) EU (Helix project)***

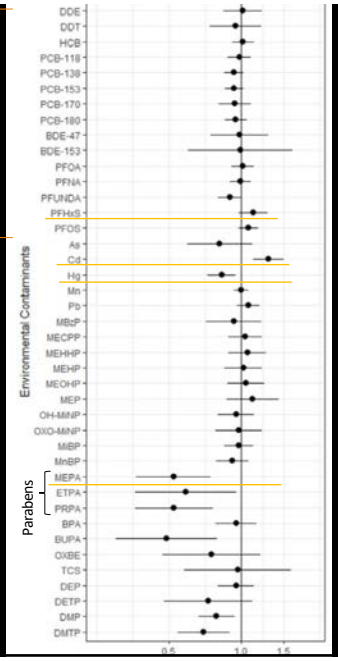
Mean pregnancy levels according to maternal education (n=1301)

Exposure change in lower Education women (ref: higher education)

Persistent organic pollutants

Metals

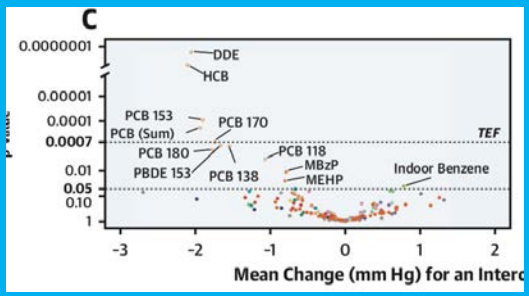
Non-persistent organic pollutants



(Montazeri, *Int J Hyg Env Heal*, 2019)

26

C. Two-layer problem: exposome-health studies



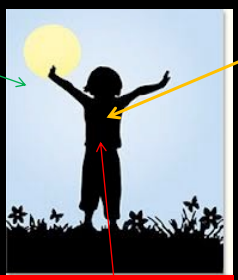
27

Assessment of the exposome of European children
Helix early-life exposome project (Vrijheid, *EHP*, 2014, Maitre, *BMJ Open*, 2018)

- Aim: to describe the early-life exposome and characterise its impact on specific health outcomes in childhood.
- Design: assessment of a wide range of external, internal exposures and 'omics markers in **1300 children from 6 European countries**.

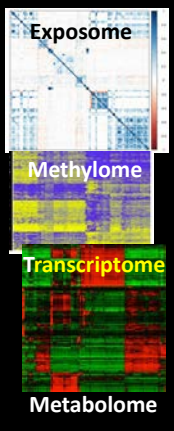
External exposome

Air pollutants (LUR models)
Passive smoking
Water pollutants
Greenspace exposure (GIS data)
Noise
UV radiation
Diet
Temperature



'omics markers

Methylome (Infinium 450k chip)
Transcriptome (mRNA, miRNA)
Metabolome (Lau, *BMC Med*, 2018)



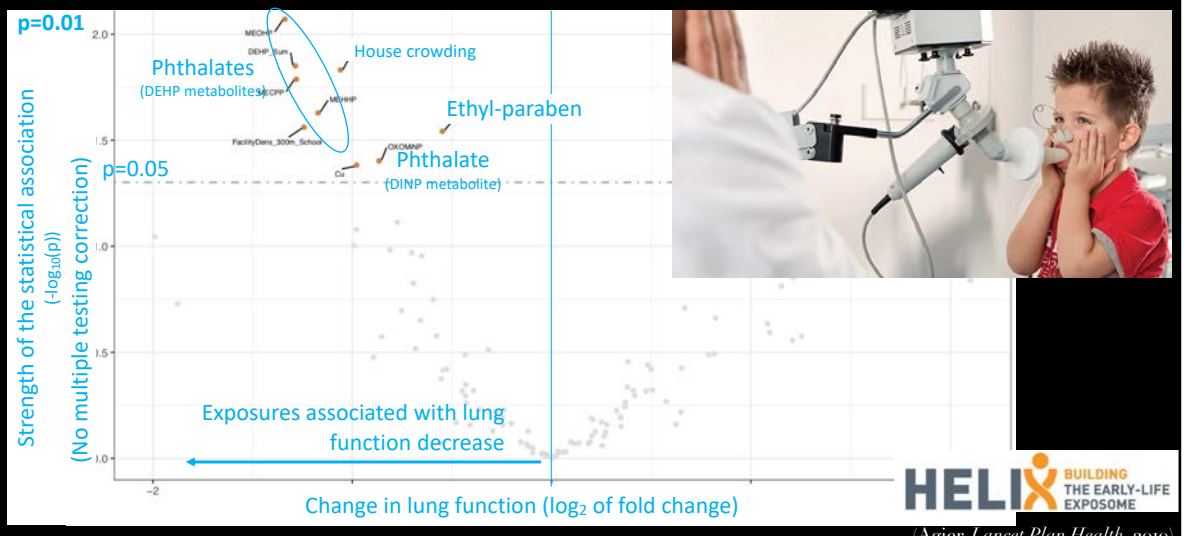
Internal exposome

Phenols, phthalates, Persistent Organic Pollutants, Heavy metals, organophospate pesticides, perfluorinated compounds…
(Haug, *Env Int*, 2018)

Health (growth, neuro-development, respiratory…)

28

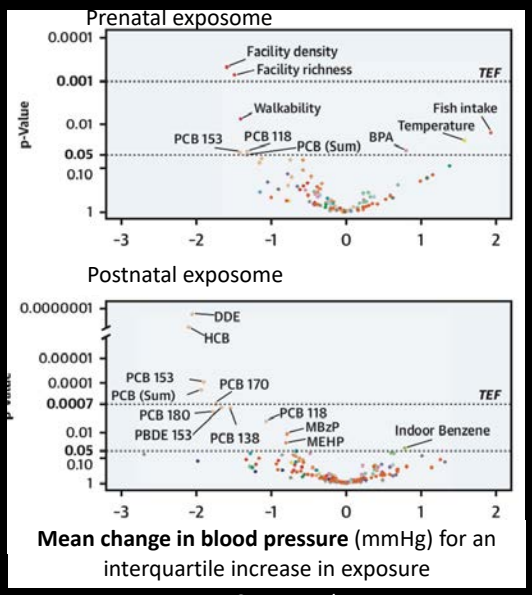
Linking the child *postnatal* exposome (125 exposures) with children lung function (FEV1 – Forced Expiratory Volume in 1s; 1033 children)



(Agier, *Lancet Plan Health*, 2019)

30

Linking the exposome with child blood pressure



Mean change in blood pressure (mmHg) for an interquartile increase in exposure

EWAS approach

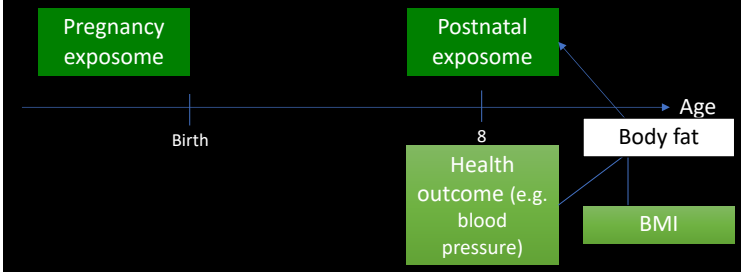
Exposures	Interquartile Range	Systolic Blood Pressure: Frequency (%) of Selection†	Systolic Blood Pressure: Adjusted Beta (95% CI)‡	Systolic Blood Pressure: p Value
Pregnancy period				
Facility density (300 m)	49.5 U/km²	98	−1.7 (−2.5 to −0.8)	0.0003
Polychlorobiphenyl 118	2.8 ng/g lipids	56	−1.4 (−2.6 to −0.2)	0.0262
Fish and seafood intake	–	50	–	0.0414
<2 times/week vs. 2-4 times/week	–	–	1.0 (−0.5 to 2.5)	0.1799
>4 times/week vs. 2-4 times/week	–	–	2.0 (0.4 to 3.5)	0.0121
Temperature (average pregnancy)	5.8°C	34	1.6 (0.2 to 2.9)	0.0240
Cotinine (µg/l)	–	8	–	0.1211
18.4-50.0 vs. <18.4	–	–	−0.8 (−2.5 to 2.8)	0.3784
>50.0 vs. <18.4	–	–	1.2 (−0.3 to 2.8)	0.1138
Bisphenol-A	4.9 µg/g creatinine	2		
Childhood period				
Hexachlorobenzene	5.1 ng/g lipids	100	−1.5 (−2.4 to −0.6)	0.0018
Dichlorodiphenyldichloroethylene	34.0 ng/g lipids	18	−1.6 (−2.4 to −0.7)	0.0004
Mono benzyl phthalate	5.5 µg/g creatinine	6	−0.7 (−1.3 to −0.1)	0.0189
Perfluorooctanoate	0.8 µg/l	6	0.9 (0.1 to 1.6)	0.0213
Temperature (daily average)	10.8°C	0		
Copper	186 µg/l	0		

DSA approach

(Warembourg, *J Am Coll Cardiol*, 2019)

31

Issues related to reverse causality (lipophilic exposures)

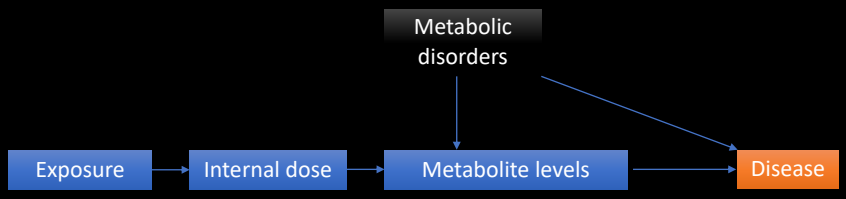


Possible cures in future studies:

- Toxicokinetic modeling
- Assay lipophilic compounds from fat biopsies
- Instrumental variables
- Increase follow-up time to limit cross-sectional analyses

32

The potential for confounding increases as you move from *external* to *molecular* exposure proxies…



(See e.g. Verner, *EHP*, 2013 or Weisskopf, *Epidemiology*, 2017)

33

Relating the exposome to health: *Methodological issues*

35

(Some) qualities of a useful (statistical) model

"I swear that the evidence that I shall give, shall be the truth, the whole truth and nothing but the truth"

High sensitivity

Low False Detection Rate (FDR)

36

4 curses,

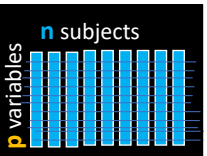
- The 2 **curses of dimension**
- **Correlation** curse
- **Mismeasurement** curse

one dream (of some)

Tackling **synergistic effects** of mixtures

37

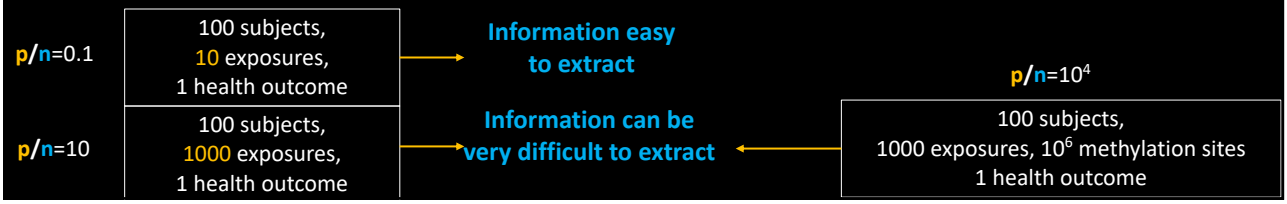
The curse of dimensionality (data sciences)



- The curse of dimensionality is a phenomenon described in data sciences according to which collecting more variables (increasing the dimension p of the data) makes statistical inference more difficult

High **p/n** ratio ➔Limited ability to detect signals

- It is related to the fact that, as the dimensionality (e.g., the number of potential predictors p of a health outcome) increases, the volume of the space increases so fast that the available data become *sparse* to an extent that can make statistical inference difficult
- In other words, an increase in the amount of data collected (*in terms of number of variables assessed p*) does not automatically translate in an increase in the extractable information, and may even lead to a decrease in information. *Data* is not *information*

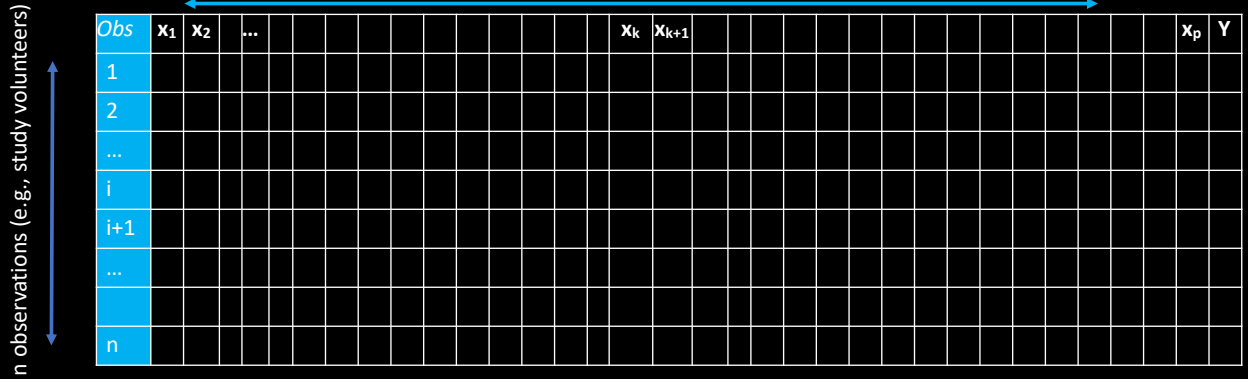


38

What happens when the dimension of a data set increases?

The 2 curses of dimension

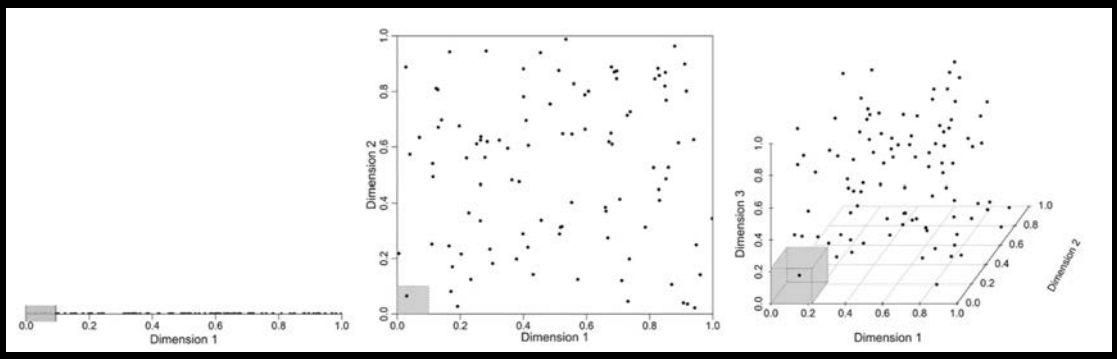
High dimension data: when the number of **variables** p becomes large compared to n



39

1. The "real" curse of dimension: *data sparsity*

The 2 curses of dimension



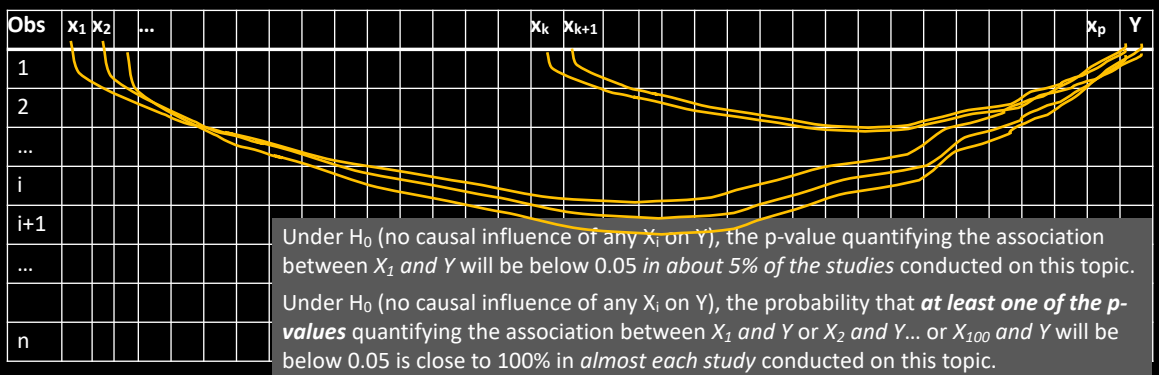
Consequences: bias; trend towards increased false negative signals (increased "type 2" error: low statistical power)

(Debie, 2017)
Cf. Stéphane Mallat, Collège de France, 2017-2018 lectures

40

2. Increased false positive rate due to multiple testing

- Remember that in exposome studies, the general aim is not a predictive one (can we predict Y given X_1, … X_p ?) but rather a causal inference aim (which exposures among X_1, … X_p causally influence Y?)



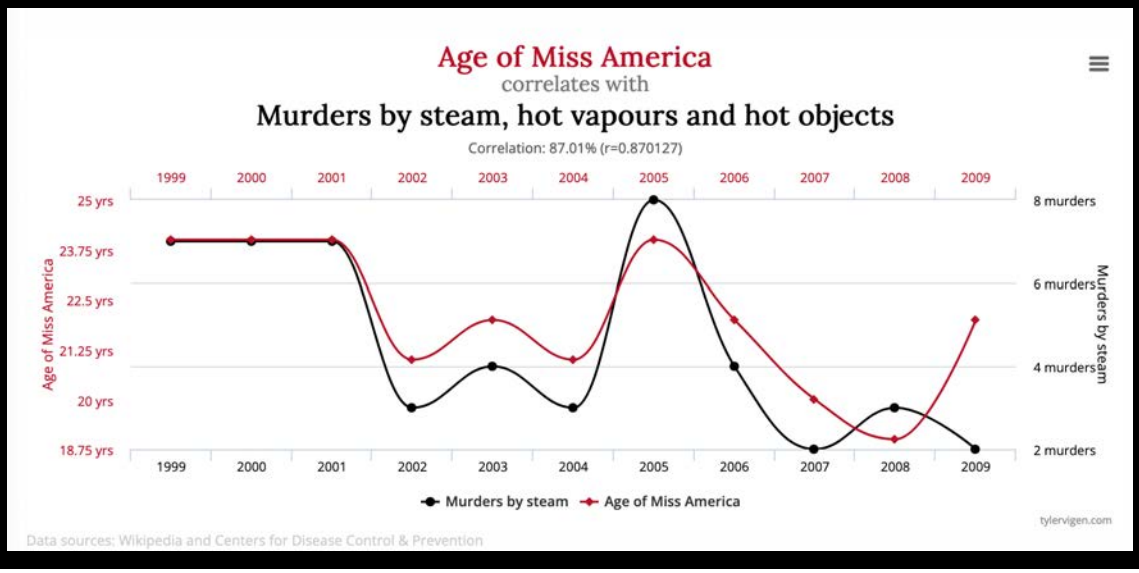
Under H_0 (no causal influence of any X_i on Y), the p-value quantifying the association between *X_1 and Y* will be below 0.05 *in about 5% of the studies* conducted on this topic.

Under H_0 (no causal influence of any X_i on Y), the probability that ***at least one of the p-values*** quantifying the association between *X_1 and Y* or *X_2 and Y*… or *X_{100} and Y* will be below 0.05 is close to 100% in *almost each study* conducted on this topic.

Consequences: increased false positive signals (increased "type 1" error)

41

An illustration of multiple testing issues



42

Background: *A very short history of multiple testing issues in social and health sciences (1)*

Quételet
(1797-1874)

- Quételet was probably the first to bring statistics from astronomy and meteorology into social and health sciences ("Research on population births, deaths, prisons, poor houses etc. in the kingdom of the Low Countries", 1827)
- Cournot-Quételet debate (1843)
  If one tests the sex ratio of all 86 French "départements" for a difference with the others, then the resulting p-values are meaningless
- In its early developments (1850-1980), epidemiology focused on "low dimension" problems with strong a priori hypotheses in which multiple testing was (at least apparently) not an issue
  John Snow and the propagation of cholera, smoking and lung cancer…

Cournot
(1801-1877)

43

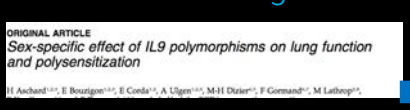
A very short history of multiple testing issues in social and health sciences (2) – The Genome era

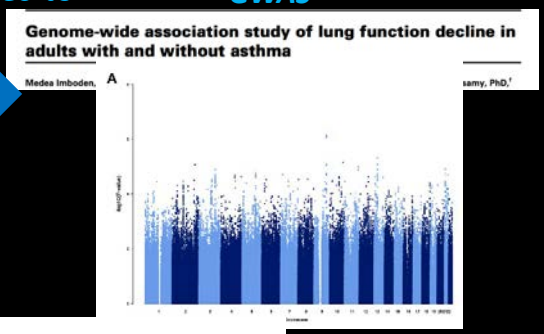
- 1950s: First multiple comparison procedures
- 1979: Holm-Bonferroni method – Family Wise Error Rates (FWER)
  Control the probability that NONE of the multiple observed scores are below a threshold α
- 1995: False Detection Rate (FDR) procedures
  Aims to make sure that the the *overall* rate of false positive signals remains below a threshold α. (Benjamini & Hochberg, *J Roy Stat Soc B*, 1995) for independent tests, Benjamini & Yekutieli (*Ann Stat*, 2001) under arbitrary dependence
- 1998: 1st commercial Affymetrix array (1494 single nucleotide polymorphisms SNPs)
- 2016: Genomic arrays feature 1.8 million genetic markers
- These multiple comparison procedures are now widely used in the genomic (and "epigenomic") literature

44

Genetics

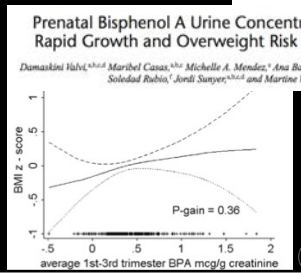
From *candidate gene studies* to **GWAS**


ORIGINAL ARTICLE
Sex-specific effect of IL9 polymorphisms on lung function and polysensitization

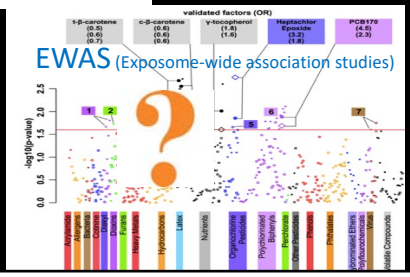

Genome-wide association study of lung function decline in adults with and without asthma

Medea Imboden, A

Environmental epidemiology


Prenatal Bisphenol A Urine Concentrations and Early Rapid Growth and Overweight Risk in the Offspring

(Valvi, *Epidemiology*, 2013)

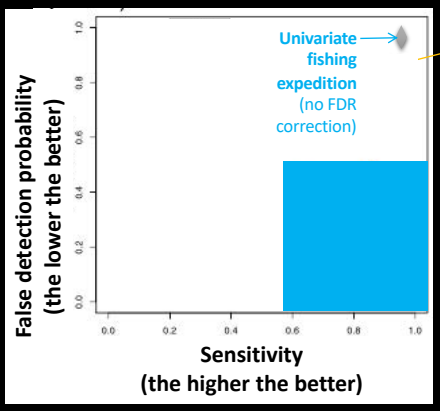


(Patel, *IJE*, 2010)

45

Efficiency of various statistical methods to relate the exposome to a health outcome
(simulation study based on realistic hypotheses)

The GWAS approach used in genetic research cannot be applied in a straightforward way (EWAS) to the exposome **(correlation curse)**



EWAS: Exposome wide association study (similar to *GWAS* approach with FDR correction)

Simulation study aiming at identifying k=1, 2, 10 or 25 real predictors out of **238** exposures (average results) among **1200** study participants

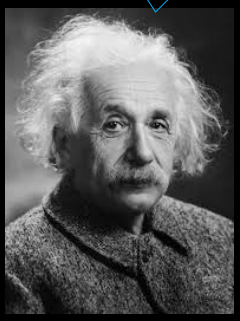
(Agier, *EHP*, 2016)

46

When a model does not fit reality…

… change reality!

… change model

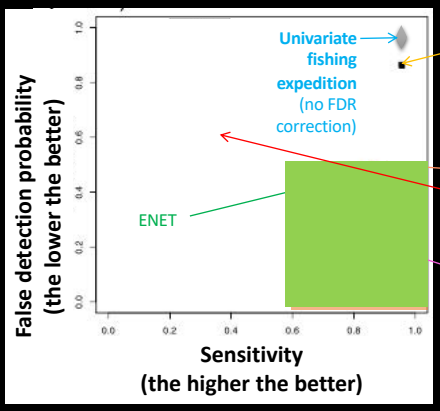


47

Efficiency of various statistical methods to relate the exposome to a health outcome

(simulation study based on realistic hypotheses)

The GWAS approach used in genetic research cannot be applied in a straightforward way (EWAS) to the exposome **(correlation curse)**



EWAS: Exposome wide association study (similar to *GWAS* approach with FDR correction)

GUESS (Bayesian variable selection method)

EWAS-MLR: EWAS followed by multiple linear regression

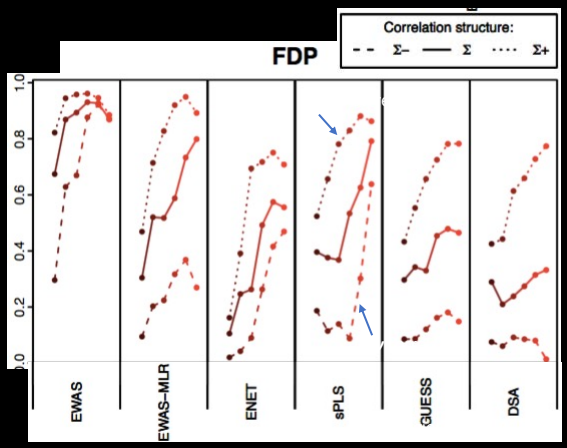
DSA (Deletion/Substitution/Addition Algorithm)

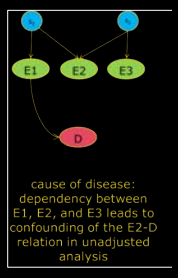
Simulation study aiming at identifying k=1, 2, 10 or 25 real predictors out of **238** exposures (average results) among **1200** study participants

(Agier, *EHP*, 2016)

48

What explains the high FDR of the EWAS approach?





cause of disease: dependency between E1, E2, and E3 leads to confounding of the E2-D relation in unadjusted analysis

False detection Proportion, (EWAS, Bonferroni correction)

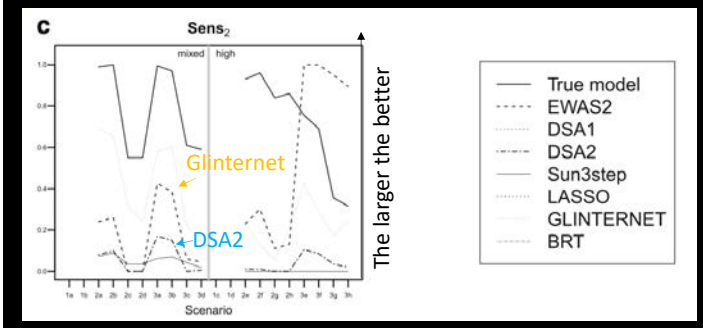
Nb of true predictors	Correlation between exposures	
	No correlation	Correlation
0	5%	2%
1	2%	65%
2	1%	81%

(Portengen et al., *unpublished results*)

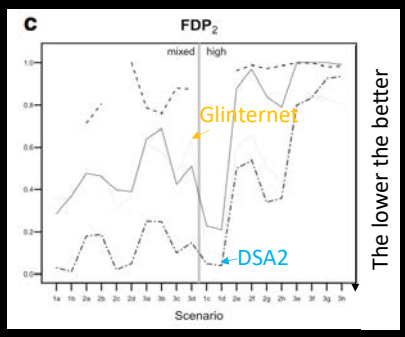
49

Are there efficient approaches to detect *interactions** between exposures?

Sensitivity to detect order-2 interaction terms



False positive rate for interaction terms



*Defined as a departure from additive effects of 2 exposures

(Barrera-Gomez, *Env Health*, 2017)

52

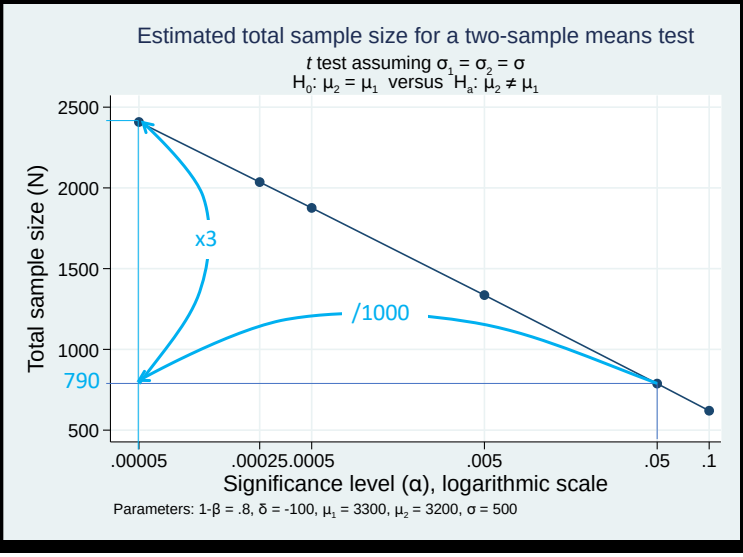
The curse of sample size: *possible answers*

1) **Increase sample size** (without increasing measurement error!)
2) **Reduce dimension**: Borrow information from toxicology (external data) or biological intermediary layers (internal data)

53

Issues related to power and sample size

Sample size required to achieve a statistical power of 80% to detect a change by 100 for a continuous outcome



If 1000 (Bonferroni-corrected) tests are performed instead of one:

sample size has to be multiplied by 3 for power to remain at 80%.

***Or* effect size has to increase by 80%**

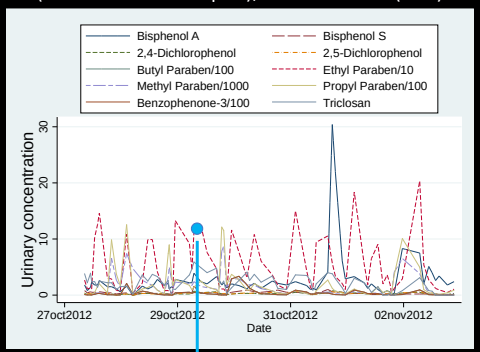
If sample and effect sizes remain constant, then power decreases to 10%.

(Note that these estimates assume a lack of correlation among the predictors, which may not be realistic)

54

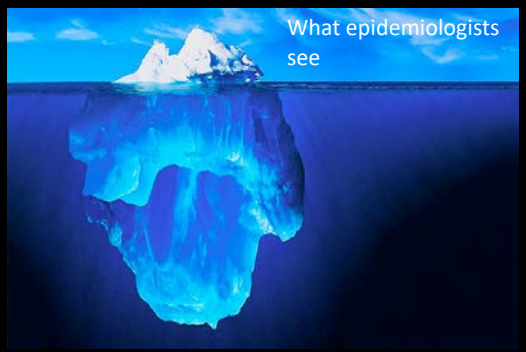
The curse of measurement error

SEPAGES-feasibility study, sampling of all urine samples for 1 woman during a week (about 70 urine samples); A. Calafat's lab (CDC)



(Vernet, *EHP*, 2018)

…leading to attenuation bias in dose-response functions

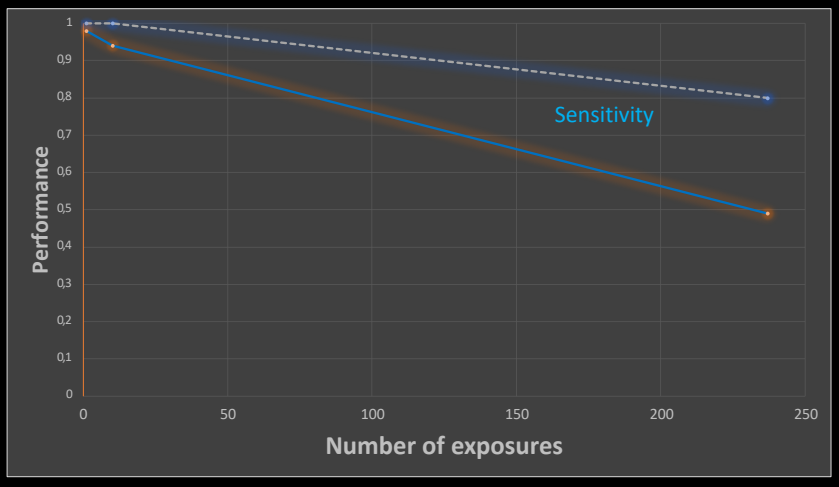


What may be the true effect

(Perrier, *Epidemiology*, 2016; Vernet, *Epidemiology*, 2019)

55

Impact of measurement error in exposome studies



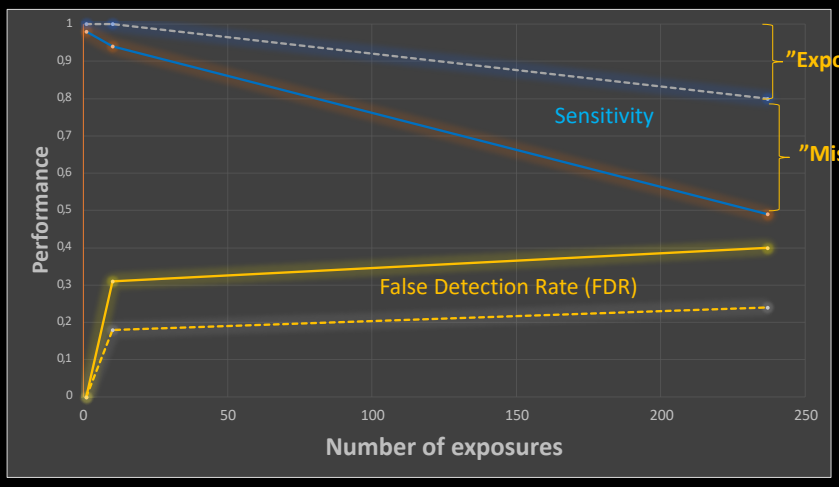
No measurement error (T_j)

Error-prone exposures (X_j)

(Agier, *Env Res*, 2020)

56

Impact of measurement error in exposome studies



"Exposome (correlation) cost"

No measurement error (T_j)

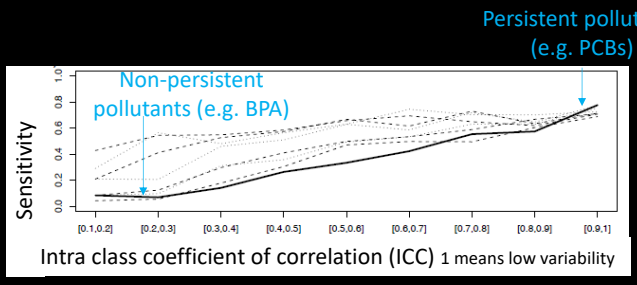
"Mismeasurement cost"

Error-prone exposures (X_j)

(Agier, *Env Res*, 2020)

57

Generalization: influence of the biomarker's variability on the sensitivity of ***exposome*** studies



The higher the temporal (within-subject) variability of a compound (low ICC), the lower the sensitivity of an exposome study to detect it.

Simulation study assuming 1200 participants and similar effect sizes for true exposures whatever their ICC.

(Agier, *Env Res*, 2020)

58

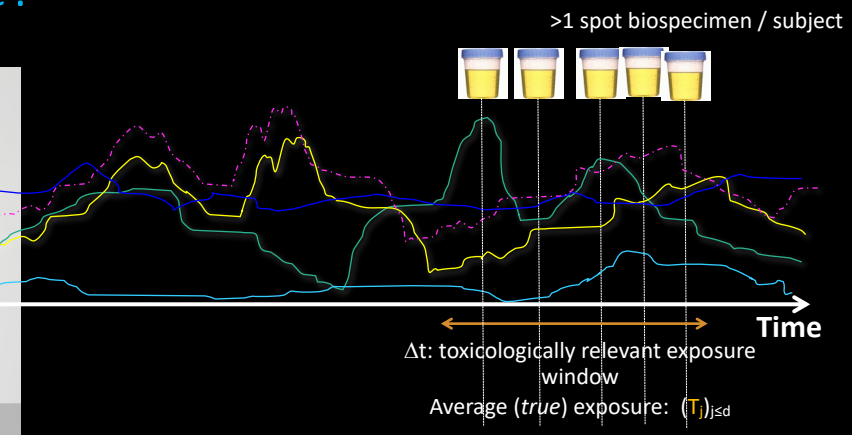
Statistical power =

f(sample size (or number of cases), exposure distribution, measurement error…)

Related to the within-subject variability of the compound

59

Is there a cure?



"Within-subject biospecimens pooling approach"

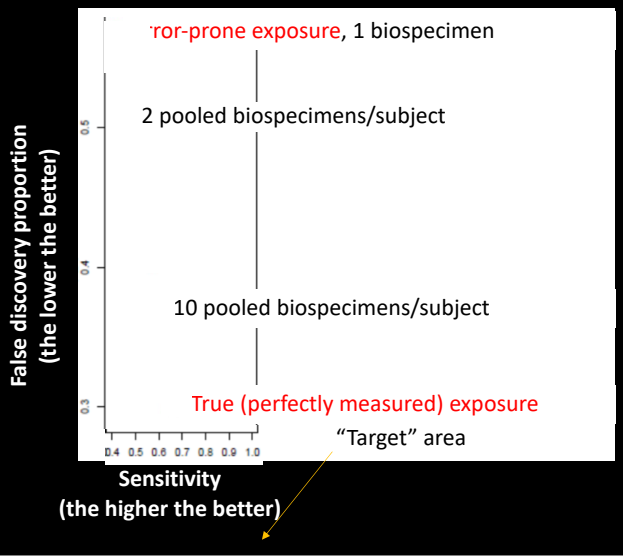
Validated in the single-exposure case (Perrier, *Epidemiology*, 2016)

Also considered in an exposome context (Agier, *Env Res*, 2020)

60

Theoretical efficiency of exposome studies relying on repeated biospecimens (simulation)

Assumption: 10 true predictors truly affects the health outcome, out of 237 exposures



(Agier, *Env Res*, 2020)

61

How many (pools) of urine samples are enough to accurately assess exposure to non persistent compounds?

Estimation of the number of urine pools needed to limit exposure mis-classification of phthalate metabolites, phenols, OP pesticide metabolites, and cotinine in pregnant women and children based on the intraclass correlation coefficients

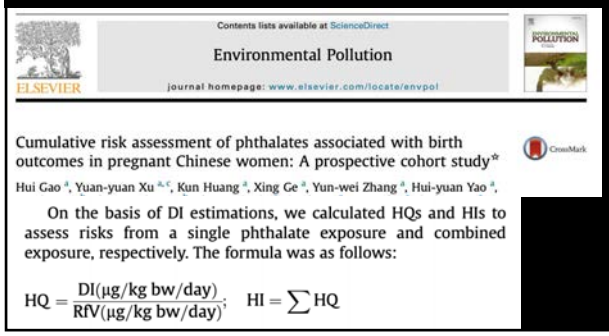
Biomarkers	Pregnant women	Children	
	N of pools of 20 urines needed for between-trimester ICC ≥ 0.80	N of pools of two urines needed for between-week ICC ≥ 0.80	N of pools of 15 urines needed for between-season ICC ≥ 0.80
Phthalate metabolites			
MEP	3	3	4
MiBP	3	3	3
MnBP	2	3	4
MBzP	3	3	3
MEHP	3	3	2
MEHHP	4	3	3
MEOHP	4	3	3
MECPP	3	3	3
oh-MiNP	4	3	4
oxo-MiNP	4	4	4
Phenols			
MEPA	4	3	4
ETPA	3	3	4
PRPA	4	3	4
BUPA	3	3	4
BPA	4	3	5
OXBE	3	1	3
TCS	3	1	5
OP pesticide metabolites			
DMP	4	4	5
DMTP	4	4	4
DMDTP	6	4	4
DEP	3	4	4
DETP	4	4	4
DEDTP	4	4	4
Cotinine	2	–	3

(Casas, *Env Int*, 2018)

62

The curse of sample size: *possible answers*

1) **Increase sample size** (without increasing measurement error!)
2) **Reduce dimension**: Borrow information from toxicology (external data) or biological intermediary layers (internal data)



Contents lists available at ScienceDirect

Environmental Pollution

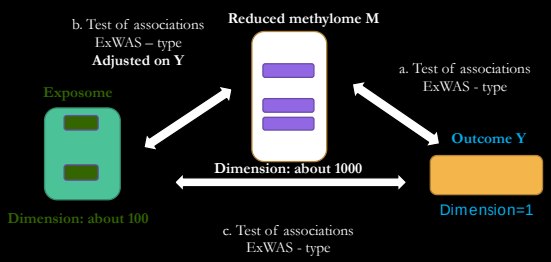
journal homepage: www.elsevier.com/locate/envpol

Cumulative risk assessment of phthalates associated with birth outcomes in pregnant Chinese women: A prospective cohort study☆

Hui Gao [a], Yuan-yuan Xu [a, c], Kun Huang [a], Xing Ge [a], Yun-wei Zhang [a], Hui-yuan Yao [a],

On the basis of DI estimations, we calculated HQs and HIs to assess risks from a single phthalate exposure and combined exposure, respectively. The formula was as follows:

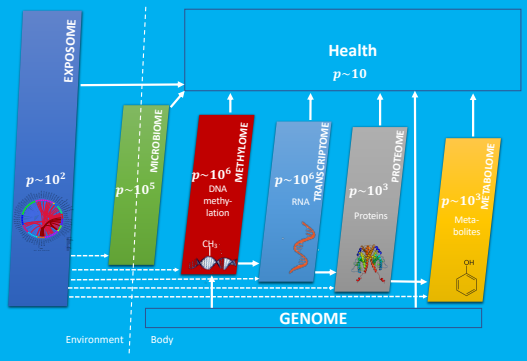
$$HQ = \frac{DI(\mu g/kg\ bw/day)}{RfV(\mu g/kg\ bw/day)}; \quad HI = \sum HQ$$



(Cadiou, *Env Int*, 2020
Cadiou, *Env Int*, 2021)

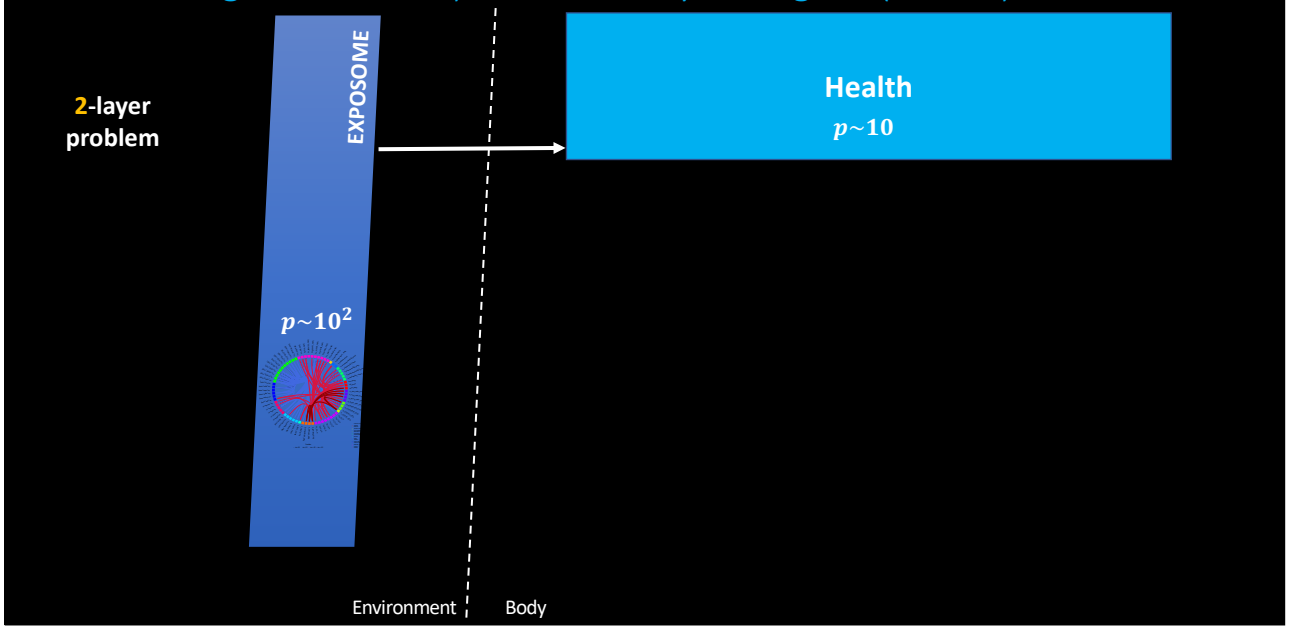
63

D. More layer: cross-omics analyses



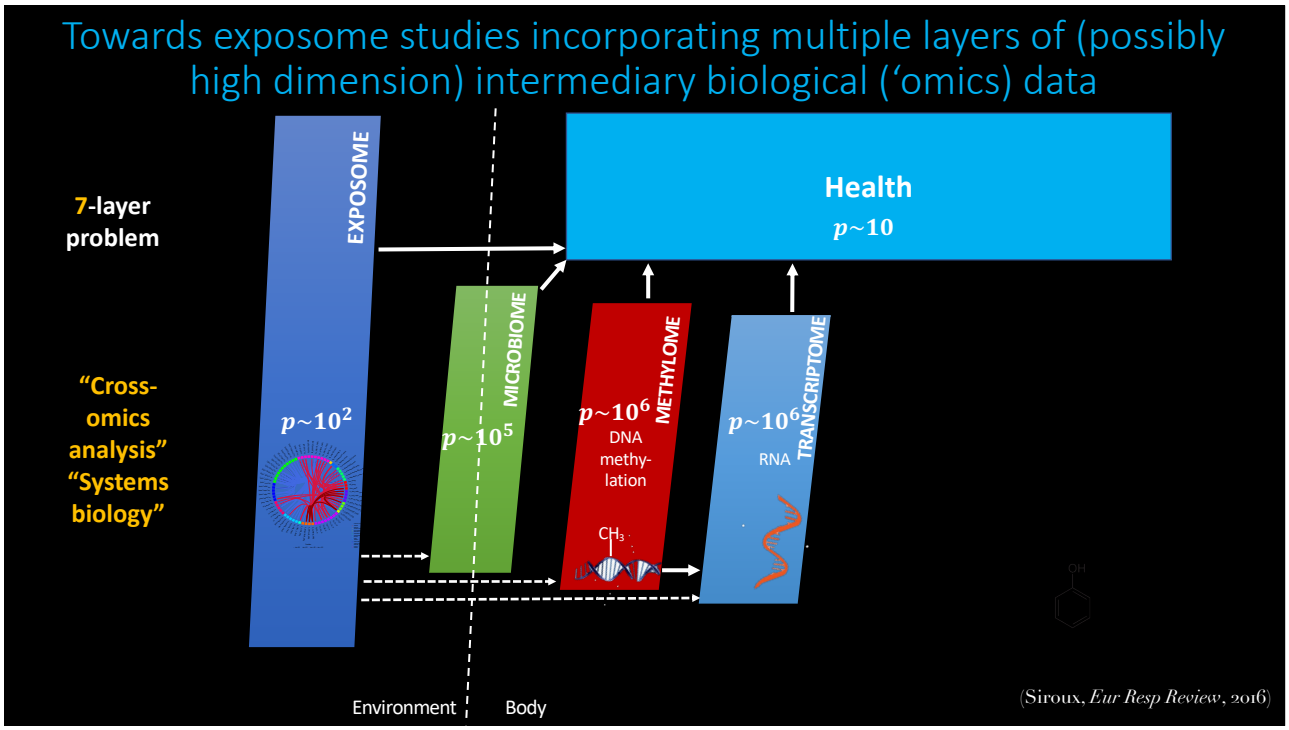
64

Towards exposome studies incorporating multiple layers of (possibly high dimension) intermediary biological ('omics) data



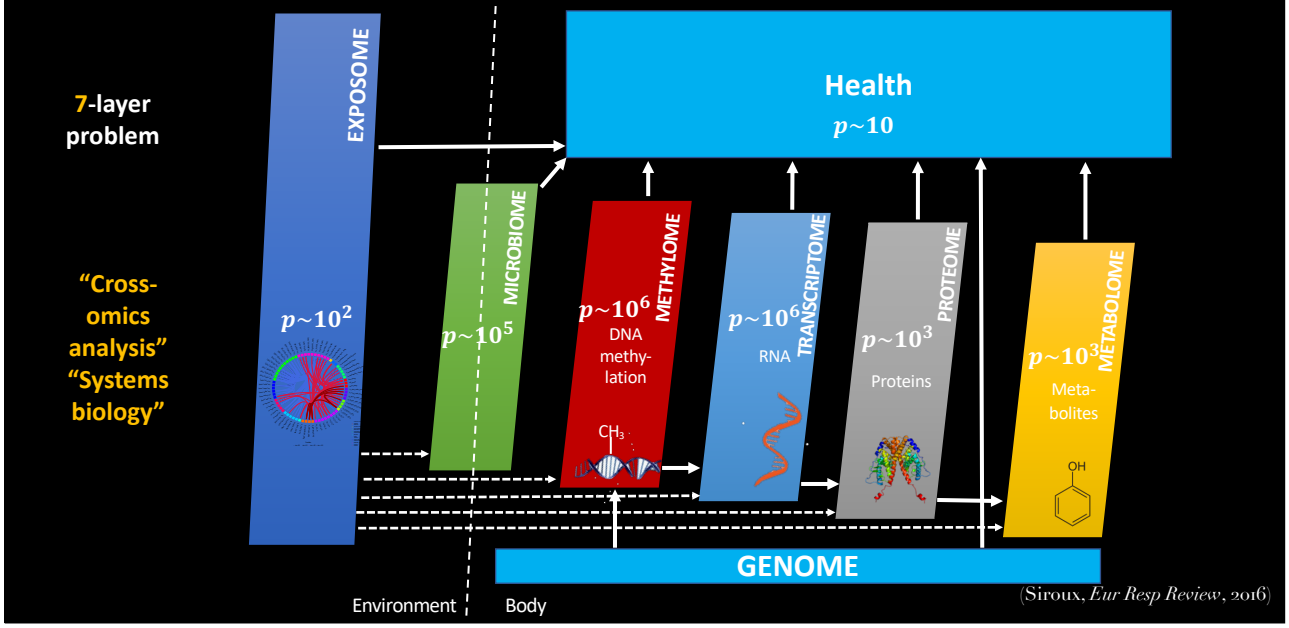
65

Towards exposome studies incorporating multiple layers of (possibly high dimension) intermediary biological ('omics) data



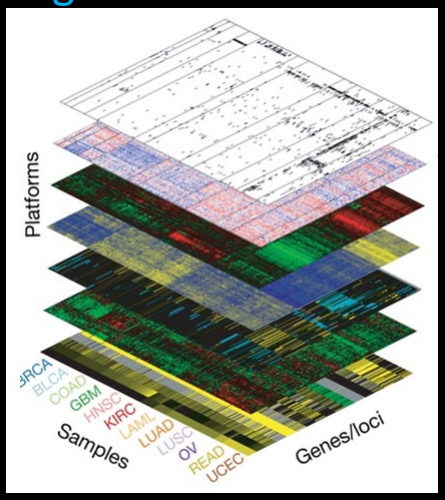
66

Towards exposome studies incorporating multiple layers of (possibly high dimension) intermediary biological ('omics) data



67

Example: HELIX project has characterized multiple layers 'Omics signatures in 1200 children



In 1200 children (6-10 years)

Platform	# features
Urine metabolome	44
Serum metabolome	177
Proteome	36
miRNAs	359
Transcriptome	35,841
Methylome	386,518
Exposome	**Ca. 120**
Health	**Ca. 10**

(Gallego-Paüls, *BMC Medicine* 2021)

68

Exposome studies incorporating multiple layers of intermediary biological ('omics) data: ***Motivation***

1. Increase the level of evidence for the exposome (or a specific set of exposures) influencing disease risk, through identification of toxicologically-plausible biological mediators
2. Adopt a "systems biology" approach, i.e. decipher the causal relations among the multiple biological layers (high dimension multilayer causal inference)

Causal inference in a multi-layer (≥3) setting *(very ambitious)*

3. Solve more easily the 2-layer problem of exposome-health relations by borrowing information from other intermediary biological layers
   - Dimension reduction
   - Discard reverse causality
   - Identify exposure biomarkers or effect biomarkers

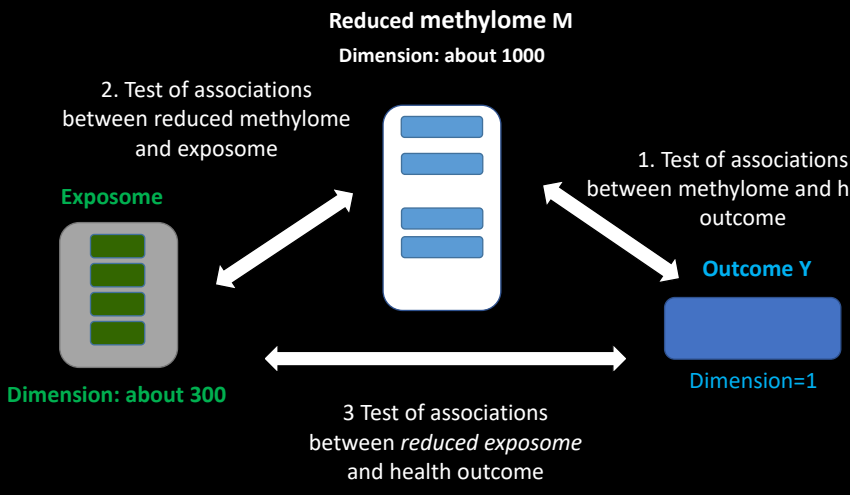
Causal inference in a 2-layer setting *(ambitious but realistic)*

4. Disease **risk prediction** (without necessarily pretention to causal inference)

Risk prediction *(technically realistic?)*

69

Solving more easily the 2-layer problem of exposome-health relations by borrowing information from another intermediary biological layer

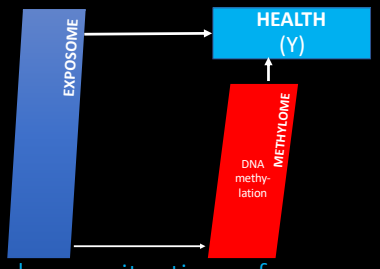


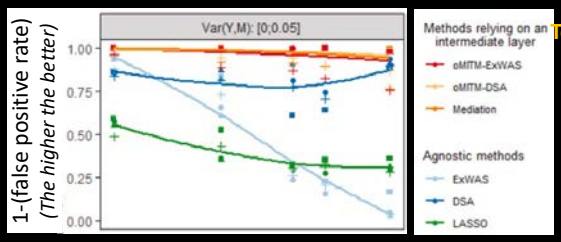
(Cadiou, *Env Int*, 2020
Cadiou, *Env Int*, 2021)

70

Taking into account information on the methylome can help unravel information on the exposome-health relation…

Hypothesis: **Causal effect** of the exposome on health

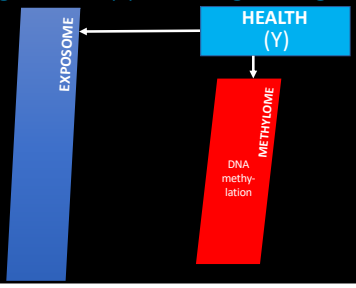


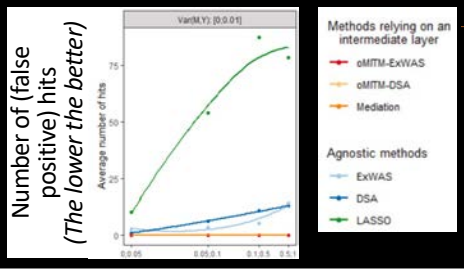


Tend to perform better

…and discard some situations of reverse causality, what an agnostic approach ignoring the methylome cannot do

Hypothesis: **Reverse causality**





Tend to perform better

(Cadiou, *Env Int*, 2021)

71

Exposome studies incorporating multiple layers of intermediary biological ('omics) data: *Motivation*

1. Increase the level of evidence for the exposome (or a specific set of exposures) influencing disease risk, through identification of toxicologically-plausible biological mediators
2. Adopt a "systems biology" approach, i.e. decipher the causal relations among the multiple biological layers (high dimension multilayer causal inference)
3. Solve more easily the 2-layer problem of exposome-health relations by borrowing information from other intermediary biological layers
   - Dimension reduction
   - Discard reverse causality
   - Identify exposure biomarkers or effect biomarkers
4. Disease **risk prediction** (without necessarily pretention to causal inference)

Causal inference in a multi-layer (≥3) setting ***(very ambitious)***

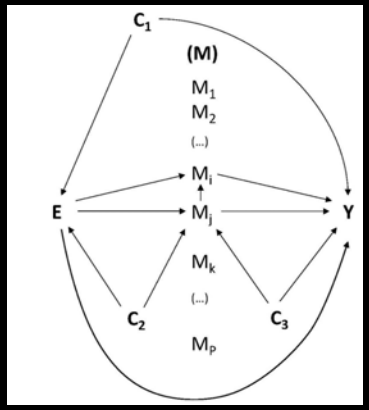
Causal inference in a 2-layer setting ***(ambitious but realistic)***

Risk prediction ***(technically realistic?)***

73

High-dimension mediation analysis

- Mediation analysis is conceptually well framed with a mediator of dimension 1 (VanderWeele, *Oxford Univ Press*, 2015). The theoretical framework in particular implies that the causal model is known a priori.
- Such an assumption is not very realistic if the mediator has a high dimension and is treated as a large set of potential mediators
  - Remember that *statistically*, models generally do not allow to infer the direction of any causal effect between A and B (i.e., A->B and B->A are not distinguishable by pure statistical tools)
  - And that biologically there may be complex causal relations within a biological layer (e.g., the methylome)
- Consequently, although the literature is full of examples of high dimension analysis relying e.g., on methylome or metabolome data, rigorously identifying the "causal" mediators or the share of the effect of E on Y mediated by an intermediary biological layer generally remains a challenge.



VanderWeele, *Epid Meth*, 2014; Blum, *EHP*, 2020)

74

Exposome studies incorporating multiple layers of intermediary biological ('omics) data: *Motivation*

1. Increase the level of evidence for the exposome (or a specific set of exposures) influencing disease risk, through identification of toxicologically-plausible biological mediators
2. Adopt a "systems biology" approach, i.e. decipher the causal relations among the multiple biological layers (high dimension multilayer causal inference)

Causal inference in a multi-layer (≥3) setting ***(very ambitious)***

3. Solve more easily the 2-layer problem of exposome-health relations by borrowing information from other intermediary biological layers
   - Dimension reduction
   - Discard reverse causality
   - Identify exposure biomarkers or effect biomarkers

Causal inference in a 2-layer setting ***(ambitious but realistic)***

4. Disease **risk prediction** (without necessarily claim to causal inference)

Risk prediction ***(technically realistic?)***

75

Disease risk prediction using '*omics* data

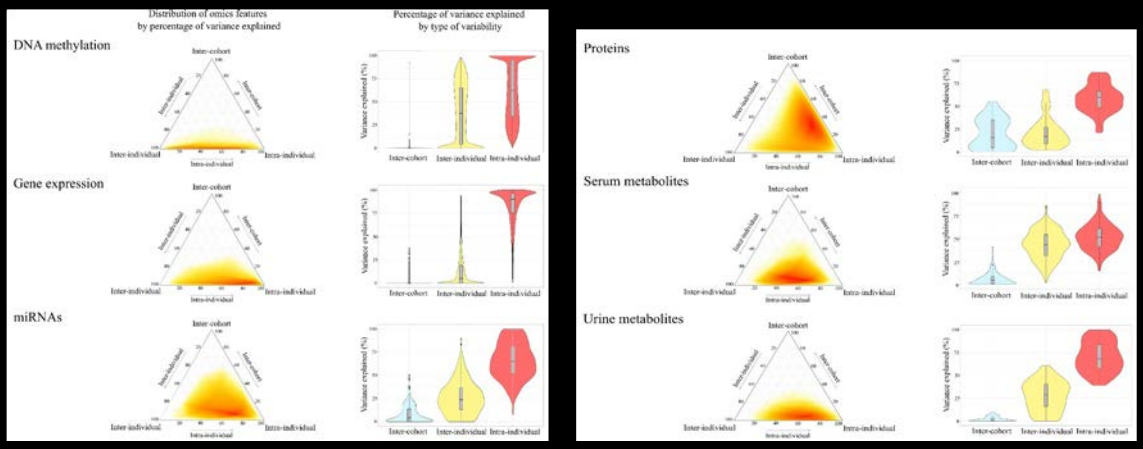
- Disease risk prediction (with claim to identifying causal health predictors) is not particularly relevant for exposome research, in which one generally aims to identify actionable environmental disease drivers likely to allow public health improvement
  - *Exception*: identification of predictors of exposures from omics (e.g., methylome) signals (Guida, *Mol Hum Gen*, 2015)
- Statistical learning tools perform generally well when it comes to prediction (as opposed to causal inference).
- However, they tend to do so when the number of "training samples" is large, which is typically not the case currently for 'omics studies in the health field, which are generally conducted on a low number of subjects (n ca. 10^3-10^4, for a number of features generally in the 10^4-10^8 range)
- Internal validation of models provides over-optimistic estimations of the classification accuracy (i.e., overfitting). Leave one out cross-validation seems particularly prone to such overconfidence in the predictive ability (Rodriguez-Perez, *Anal Bioanal Chem*, 2018)

76

Omics markers are also highly variable

- Intra-, inter-individual and cohort variability of multi-omics profiles **measured 6 months apart** in 156 children

- DNA methylation most stable; expression, least stable
- Strong heterogeneity between features



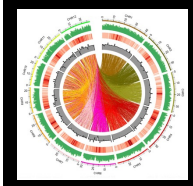
(Gallego-Paüls, *BMC Medicine* 2021)

77

E. Perspectives of exposome research

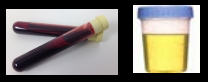
78

The historian's corner: A parallel between *genome* and *exposome* research



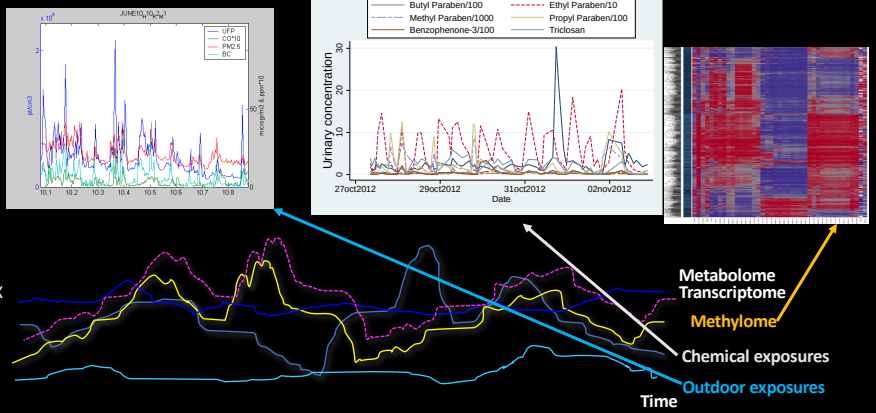
GENOME

EXPOSOME

	A scientific /technical achievement	Powerful measurement devices	Annotated biologically-relevant information units	A stable and easily accessible biological support of information	Statistical tools, epidemiological designs
GENOME	**HGP**: Sequencing of 	High-throughput sequencing platforms	Genes	 DNA	**G**WAS, multiple testing correction techniques, large consortia *Directly transferable to the exposome?* ↓ **E**WAS? (Agier, *EHP*, 2016)
EXPOSOME	**NHANES study (USA)**: 	Chromatography Mass spectrometry Coupling of both (UPLC-HRMS) + dosimeters + environmental models	Exposure biomarkers	Urine, blood (other matrices) 	

79

Continuous monitoring of multiple time-varying layers of data



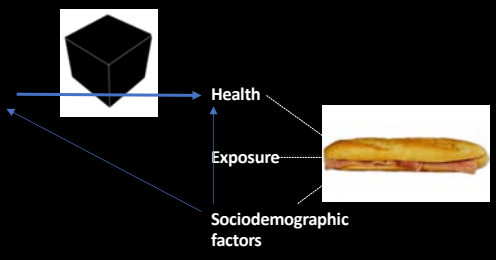
80

Cohorts of connected study participants

Enki Bilal

81

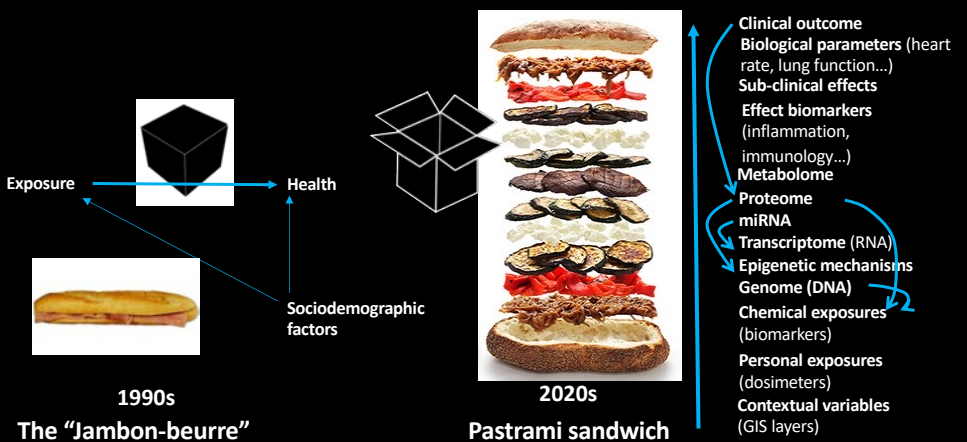
More complex causal structures



1990s
The "Jambon-beurre" (or *black box*) era

82

More complex causal structures



1990s
The "Jambon-beurre" (or *black box*) era

2020s
Pastrami sandwich (or *Chinese boxes*) era

83

Can reliance on an exposome approach be a cure to some old issues?

Could be taken care of in an exposome setting

- E_i may be measured with error (**measurement error**)
- Confounders (and outcome) may be measured with error
- There may be confounding by other unmeasured exposures (**confounding by co-exposures**) **X**
- There may be hidden **multiple testing** (e.g., because actually E_j was not an a priori choice) **(X)**
- **Random fluctuations** → **Still an (even greater) issue in exposome studies**
- **Rather *low throughput*** approach compared e.g. to toxicology (1 exposure every 1-2 years per PI in a given cohort?) **X**
- The study is more likely to be published if it highlights a "significant" association, biasing the meta-analysis (**publication bias**) **X**

Issue of false positive findings

84

Growing in number or drowning by numbers?

Not safe to increase the number of exposures considered…

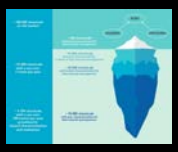
-if you cannot simultaneously improve the quality of their assessment

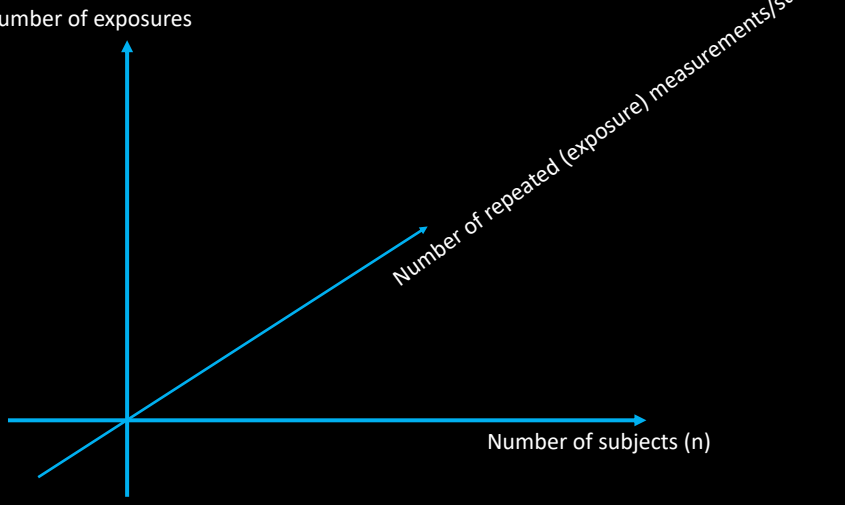
(which can be done by increasing the number of biospecimens collected per subject; see Perrier, *Epidemiology*, 2016)

-if you cannot simultaneously increase sample size

86

Optimising the design of an exposome study





87

Some possible questions

1) What happens to health outcome c if E_i varies? (e.g., : does smoking cause lung cancer?) [**classical, single exposure, aetiology**]
2) What are all the external risk factors that influence c? [exposome-health study/**exposome-wide study**]
3) What happens to health if E_i varies? (e.g., : what are the sanitary consequences of smoking?) [**outcome-wide aetiology**]
4) What intermediary variables may explain an effect of E_i on c? [**Mechanistic research**/mediation analysis]
5) Which risk factor has the largest *impact* (i.e. attributable number of disease cases) on c today? [**Environmental burden of disease**] On health overall?
6) Are specific sociodemographic subgroups disproportionately exposed to harmful exposures? [**Environmental justice problem**]

Still a challenge (power, FDP)

Feasible (but not an exposome study)

Conceptual issues

Complex

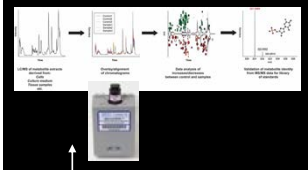
Within-reach

91

The *Exposome/environmental health* ambitious road

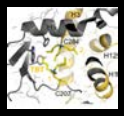
Exposome

Quantification
Platforms
TD-TK modeling
Dosimeters
« Exposome-ready » cohorts



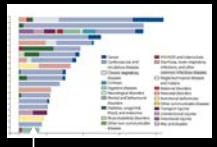
Link with socio-territorial characteristics
« Environmental justice »

Mechanisms of action
Cross-omics studies
In vitro and in vivo toxicology, cohorts



Health effects
In vivo toxicology (AOPs), Cohorts

Health impact
Environmental disease burden



Act on the exposome
Individual interventions
Governance / regulatory framework

Methodological issues

92

As a conclusion

- The exposome is a challenging but fruitful concept for environmental health research
- It allows to explicitly tackle essential challenges sometimes hidden in single exposure environmental epidemiology
  - Publication bias, differential measurement error across exposures…
- Exposome studies shouldn't consist in studies with a number of subjects is similar (or lower, as is often the case) to single exposure studies, with more exposures assessed.
  - Design issues, to some extent already faced when genetic epidemiology went genome-wide
- The (probably less challenging) outcome-wide approach (VanderWeele, *Epidemiology*, 2016) is also worth considering when it comes to improving the throughput of human studies.
- Epidemiology and toxicology need each other to walk along the way of the exposome
  - E.g., using toxicology to a priori reduce the dimension of epidemiological exposome studies

93