Estimating the architecture of exposome-phenome association and implications for precision health

Chirag J Patel

NASEM Standing Committee on diet-disease relationships

3/19/2021

HARVARD MEDICAL SCHOOL

BLAVATNIK INSTITUTE BIOMEDICAL INFORMATICS

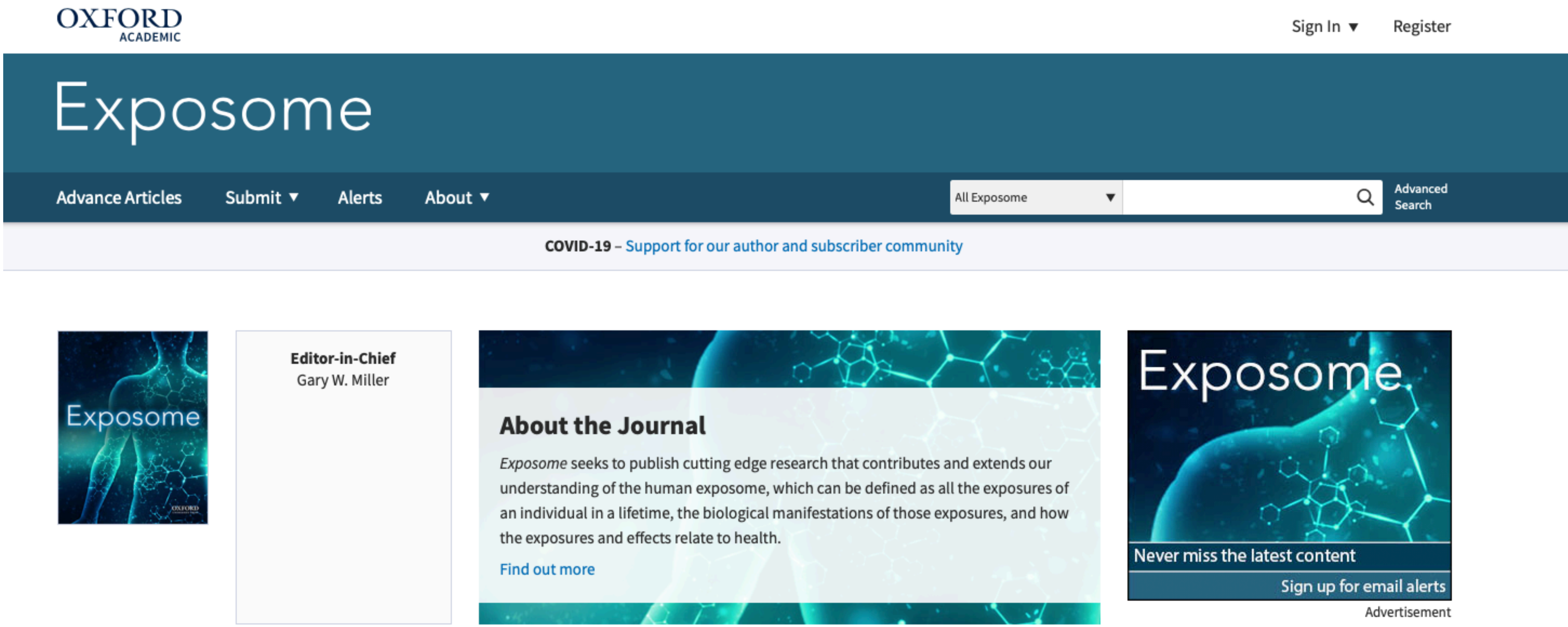
chirag@hms.harvard.edu

@chiragjp

www.chiragjpgroup.org

Conflicts of Interest

- National Institutes of Health
- National Science Foundation
- United Health Group
- Janssen
- Sanofi
- XY Health, Inc



https://academic.oup.com/exposome

“<u>*Intensive lifestyle”*</u> *over multiple domains (diet + physical activity)*, weight loss, and reduction in T2D incidence

Variable	No. of Participants (%)	Incidence			Reduction in Incidence (95% CI)*		
		Placebo	Metformin	Lifestyle	Lifestyle vs. Placebo	Metformin vs. Placebo	Lifestyle vs. Metformin
		cases/100 person-yr			percent		
Overall	3234 (100)	11.0	7.8	4.8	58 (48 to 66)	31 (17 to 43)	39 (24 to 51)
Age							
25–44 yr	1000 (30.9)	11.6	6.7	6.2	48 (27 to 63)	44 (21 to 60)	8 (−36 to 37)†
45–59 yr	1586 (49.0)	10.8	7.6	4.7	59 (44 to 70)	31 (10 to 46)	41 (18 to 57)†
≥60 yr	648 (20.0)	10.8	9.6	3.1	71 (51 to 83)	11 (−33 to 41)	69 (47 to 82)†
Sex							
Male	1043 (32.3)	12.5	8.1	4.6	65 (49 to 76)	37 (14 to 54)	46 (20 to 63)
Female	2191 (67.7)	10.3	7.6	5.0	54 (40 to 64)	28 (10 to 43)	36 (16 to 51)
Race or ethnic group							
White	1768 (54.7)	10.3	7.8	5.2	51 (35 to 63)	24 (3 to 41)	36 (14 to 52)
African American	645 (19.9)	12.4	7.1	5.1	61 (37 to 76)	44 (16 to 63)	29 (−18 to 58)
Hispanic	508 (15.7)	11.7	8.4	4.2	66 (41 to 80)	31 (−9 to 56)	51 (13 to 72)
American Indian	171 (5.3)	12.9	9.7	4.7	65 (7 to 87)	25 (−72 to 68)	52 (−35 to 83)
Asian‡	142 (4.4)	12.1	7.5	3.8	71 (24 to 89)	38 (−55 to 75)	52 (−46 to 84)
Body-mass index§							
22 to <30	1045 (32.3)	9.0	8.8	3.3	65 (46 to 77)	3 (−36 to 30)†	63 (44 to 76)†
30 to <35	995 (30.8)	8.9	7.6	3.7	61 (40 to 75)	16 (−19 to 41)†	53 (28 to 70)†
≥35	1194 (36.9)	14.3	7.0	7.3	51 (34 to 63)	53 (36 to 65)†	−4 (−47 to 26)†
Plasma glucose¶							
In the fasting state							
95–109 mg/dl‖	2174 (67.2)	6.4	5.5	2.9	55 (38 to 68)	15 (−12 to 36)†	48 (27 to 63)
110–125 mg/dl**	1060 (32.8)	22.3	12.3	8.8	63 (51 to 72)	48 (33 to 60)†	30 (6 to 48)
Two hours after an oral load							
140–153 mg/dl	1049 (32.4)	7.1	4.3	1.8	76 (58 to 86)†	41 (11 to 61)	59 (27 to 77)
154–172 mg/dl	1103 (34.1)	10.3	6.6	4.4	60 (41 to 72)†	38 (13 to 56)	34 (2 to 56)
173–199 mg/dl	1082 (33.5)	16.1	12.3	8.5	50 (33 to 63)†	26 (3 to 43)	33 (9 to 51)

*CI denotes confidence interval.

†P<0.05 for the test of heterogeneity across strata. Age, body-mass index, and plasma glucose were analyzed as continuous variables.

‡This category includes 20 Pacific Islanders.

§The eligibility criterion was a body-mass index of at least 22 for Asians and at least 24 for all other persons.

¶To convert the values for glucose to millimoles per liter, multiply by 0.05551.

‖This category includes American Indian participants who had a fasting glucose concentration that was less than 95 mg per deciliter, according to the eligibility criteria.[6]

**This category includes 54 participants with a fasting glucose concentration of 126 to 139 mg per deciliter who were enrolled before June 1997,[6] when the eligibility criteria were changed to conform to the diagnostic criteria of the American Diabetes Association, published that year.[11]

What is “lifestyle”?

DPP NEJM 2002
Tuomilehto et al NEJM 2001

We may never find the *single* exposure
(e.g., ***smoking***) that causes disease (e.g., ***lung cancer***)…

BRITISH MEDICAL JOURNAL

LONDON SATURDAY JUNE 26 1954

THE MORTALITY OF DOCTORS IN RELATION TO THEIR SMOKING HABITS

A PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

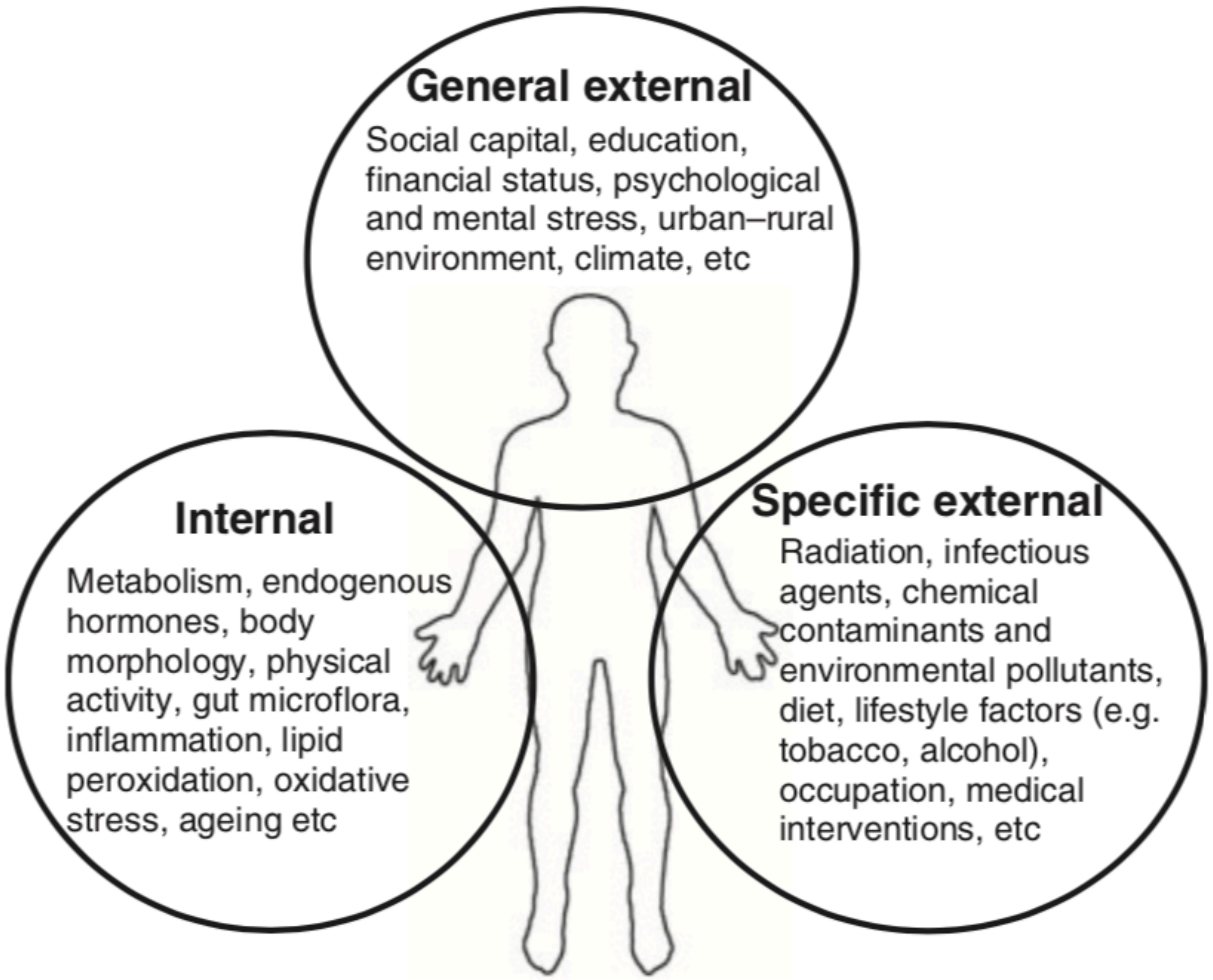
A. BRADFORD HILL, C.B.E., F.R.S.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

(Most) phenotypes/diseases are polygenic and multifactorial

Phenome	***Genome***	***Exposome***
P =	**G +**	**E**
Gene expression Type 2 Diabetes Cancer Alzheimer's	Variants	Infectious agents Nutrients Pollutants Drugs

E *Characterizing healthy lifestyles with the exposome:* the complex array of internal and external exposures humans encounter from birth to death



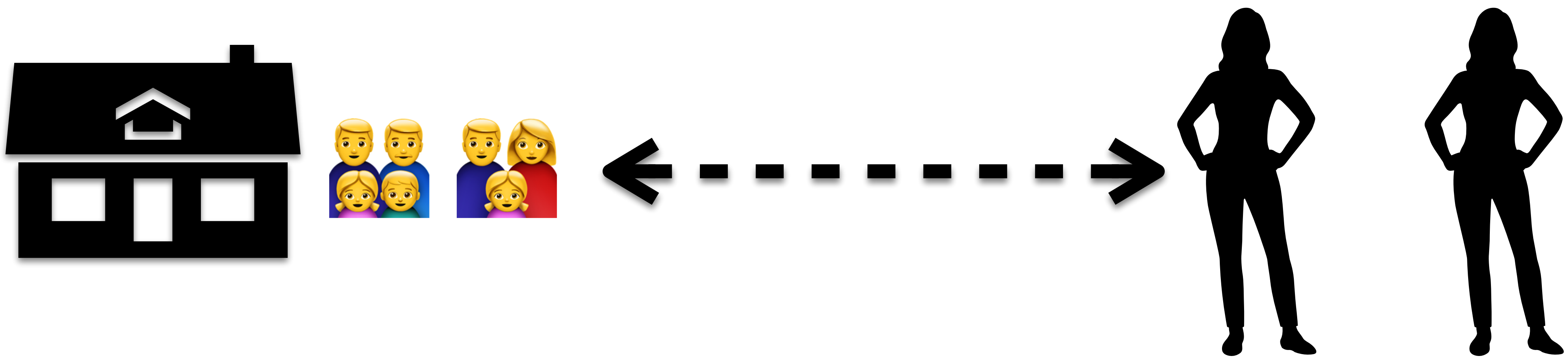
Diabetes Care 2021
Diabetologia 2021
Clinical Chemistry 2020
PLoS Comp Bio 2020
Cell Host and Microbe 2019
Nature Genetics, 2019
Aging 2019
AJE, 2015, 2019, 2019
ES&T, 2019
Environment Int, 2019
AIDS 2018
JAMA, 2014, 2018
ARPH, 2017
IJE, 2012, 2013, 2017
JCE, 2015
Proc Symp Biocomp, 2015
Reprod Tox, 2014
Hum Genet 2013
JECH, 2014
Circulation, 2012
Diabetes Care, 2012
PLoS ONE, 2010

Wild, 2005, 2012
Ioannidis, Loy, Poultin, Chia, 2009
Rappaport and Smith, 2010, 2011
Buck-Louis and Sundaram 2012
Miller and Jones, 2014
Patel and Ioannidis, 2014ab
Manrai et al 2017
Vermuelen et al 2020

The exposome is ***shared*** and ***non-shared***!

shared

non-shared



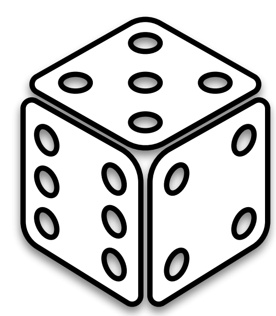
Small particles in air pollution

Differing "lifestyle" among individuals

Who is the smoker?

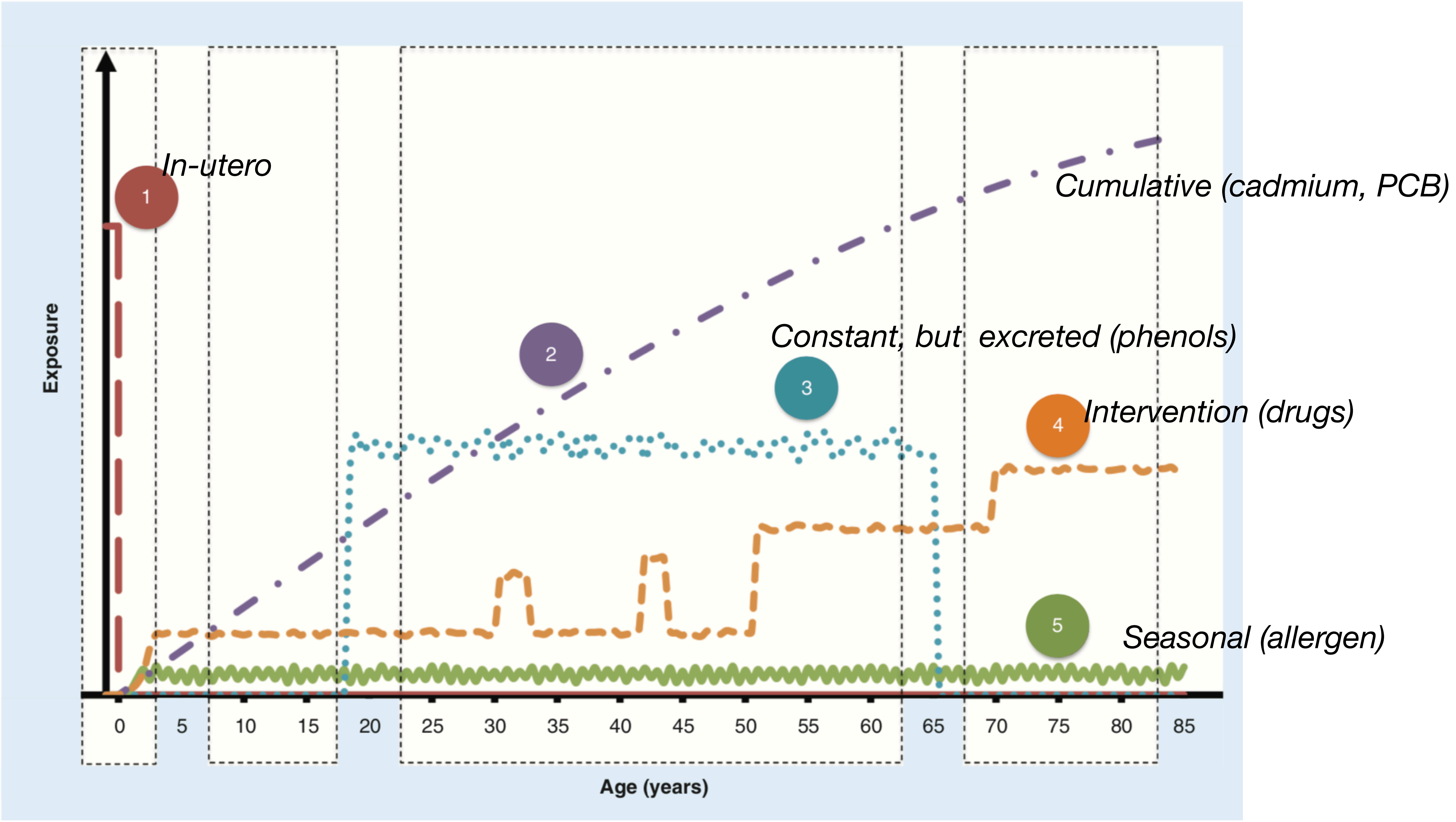
https://www.businessinsider.com/how-smoking-ages-the-face-of-identical-twins-2013-11

non-shared *E*



random chance

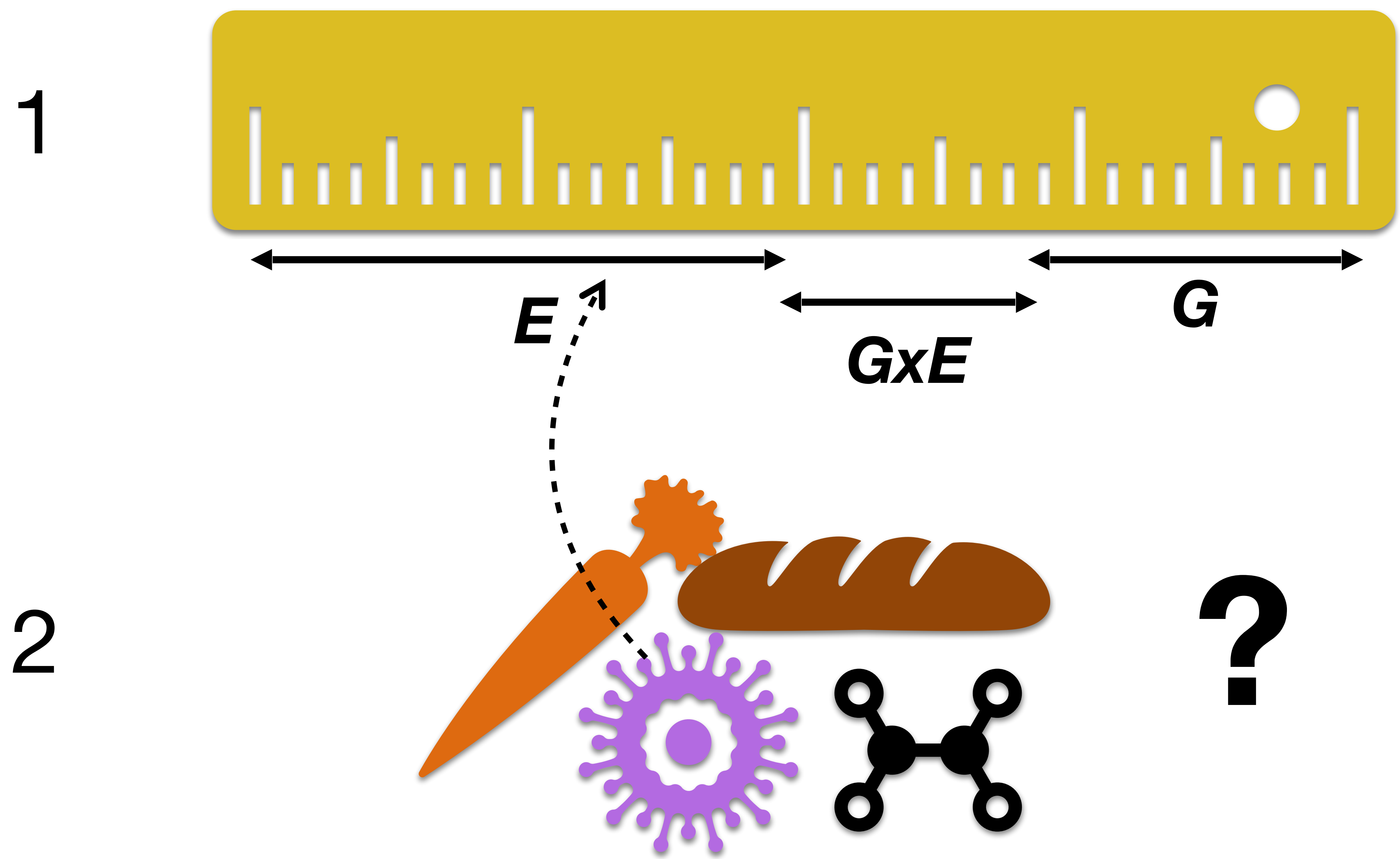
Non-shared exposome (and dietary intake) is temporally complex: *when (and at what frequency) do we measure?*



Not shown: Diurnal

Athersuch *Bioanalysis* 2012

Exposomic research for precision medicine:
Measuring ***variance explained*** and accelerating ***discovery***



$$\sigma^2_P = \sigma^2_G + \sigma^2_E$$

...

Heritability (H^2) is the range of phenotypic variability attributed to genetic variability in a population

$$H^2 = \frac{\sigma^2_G}{\sigma^2_P}$$

Indicator of the proportion of phenotypic differences attributed to **G**.

What about **E?**

$$\text{"}E^2\text{"} = \frac{?}{\sigma^2_P}$$

E^2

Combination of ***shared*** and ***non-shared*** exposome

$$\sigma^2_E = \sigma^2_{shared} + \sigma^2_{non\text{-}shared} + error$$

Shared E (C^2) is the range of phenotypic variability attributed to shared ***household*** or ***geography*** ***(but not genetics)***

$$C^2 = \frac{\sigma^2_{shared}}{\sigma^2_P}$$

Air pollution, seasonality, shared socioeconomic factors

Deconvolving** mixture of genetics, **shared E (C^2),
and indicators of shared *E (air pollution, weather, and economics) in real-world data*

Weather

Air Pollution

Census SES

+

aetna®

insurance claims

Disease (ICD9/ICD10), procedures, drugs, labs

N ~ 45M

Amassing (the largest) **twin** and ***sibling*** cohort in the US to estimate ***G*** *and* ***E*** *in* **~500 *P***

- Assume familial relationships in ***subscriber groups***
- ***Subscriber group*** less than 15 members
- Both members are child of ***primary subscriber*** (e.g., employed individual)
  - **Same date of birth**
- Year of birth ***occurs on or after 1985***
- Member ***enrollment*** greater than ***36 months***

Same Sex - Female	**17,919**
Same Sex - Male	**17,835**
Opposite Sex	**20,642**
total	**56,396**

724K siblings!

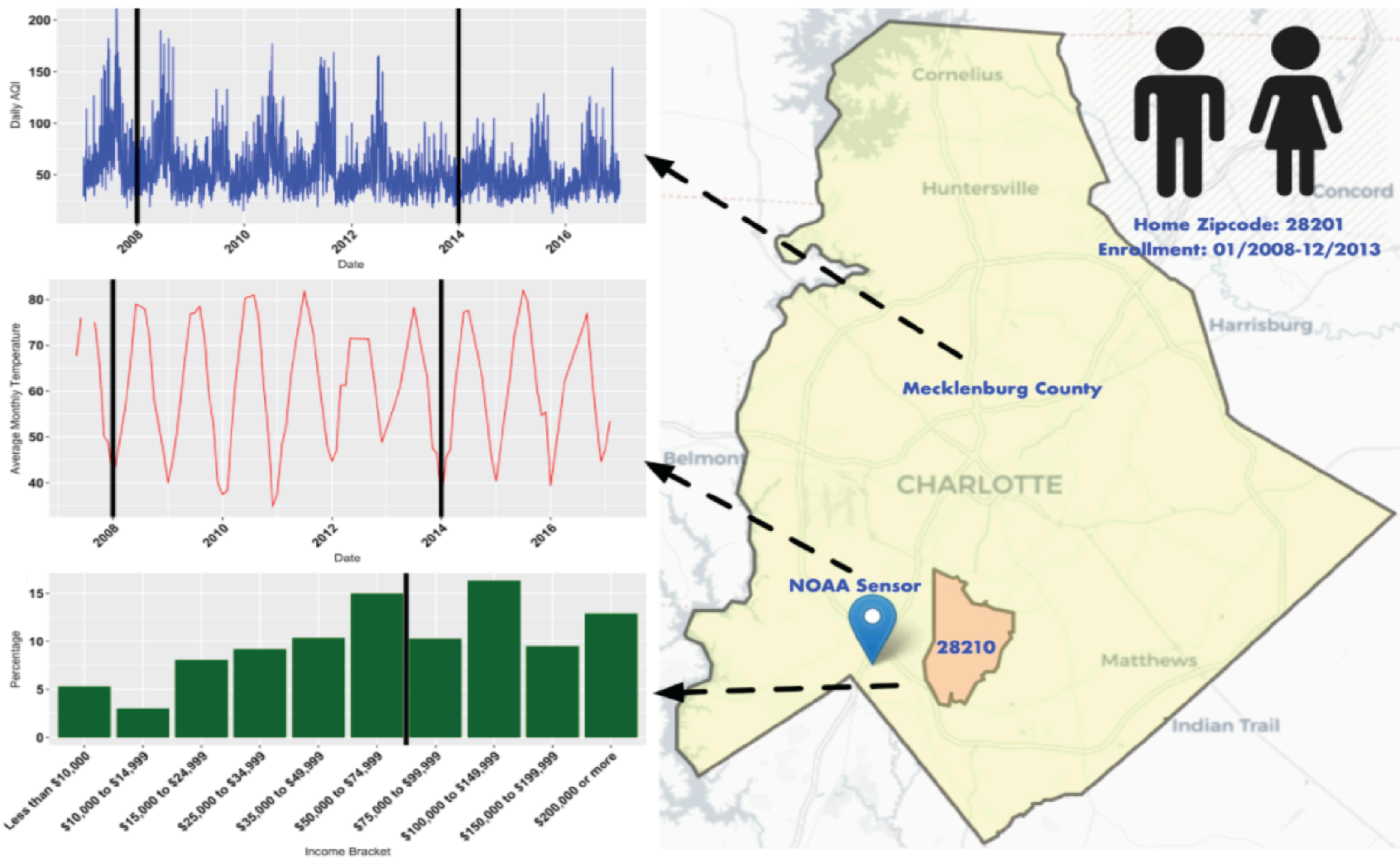
Largest collection of twins in US (next largest has ~28k pairs)

Lakhani et al., Nature Genetics 2019

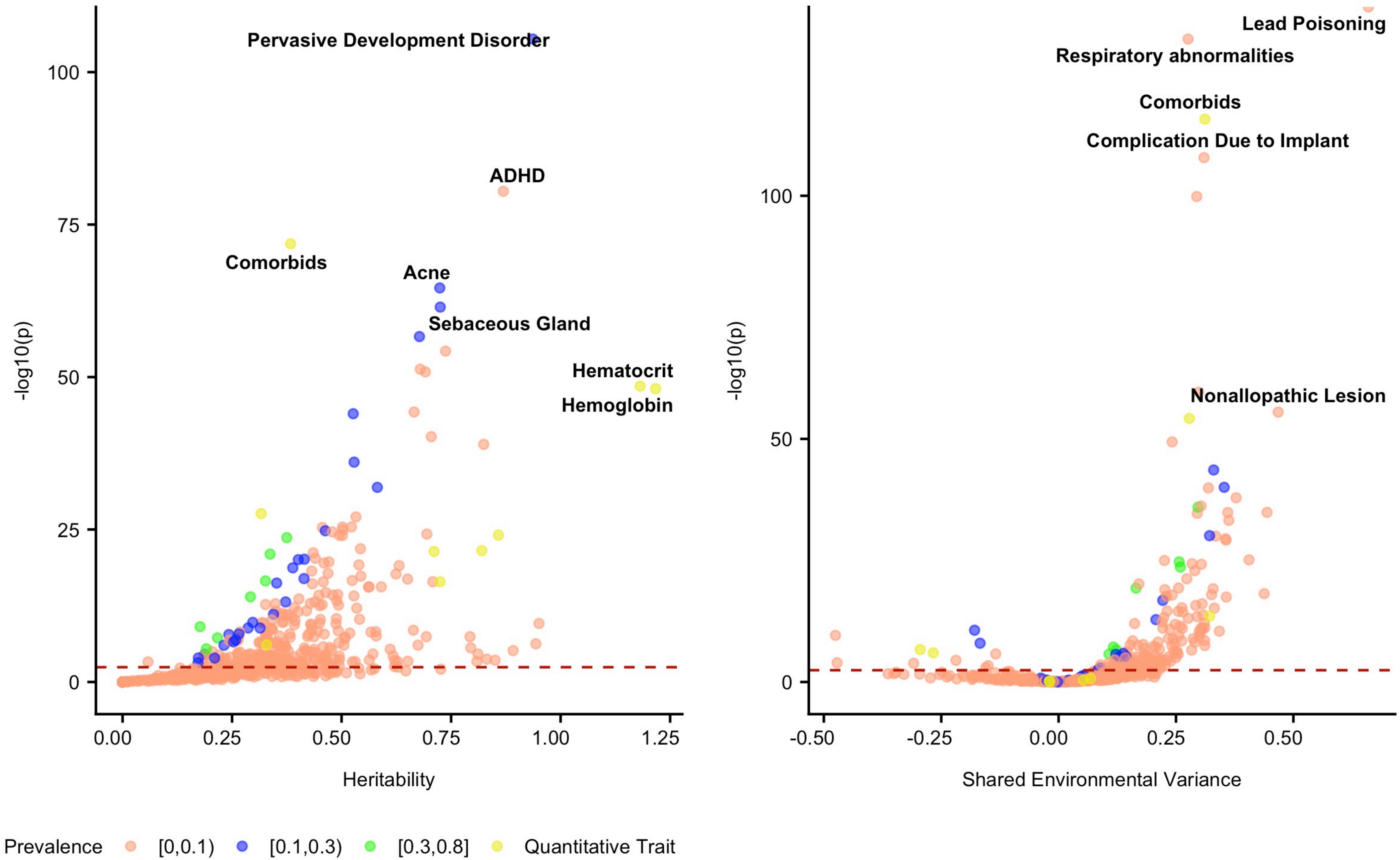
Where do we get ***E*** indicators?

Exposome Data Warehouse (~1TB)

*Geographical information system-enabled database to map individuals to **E***

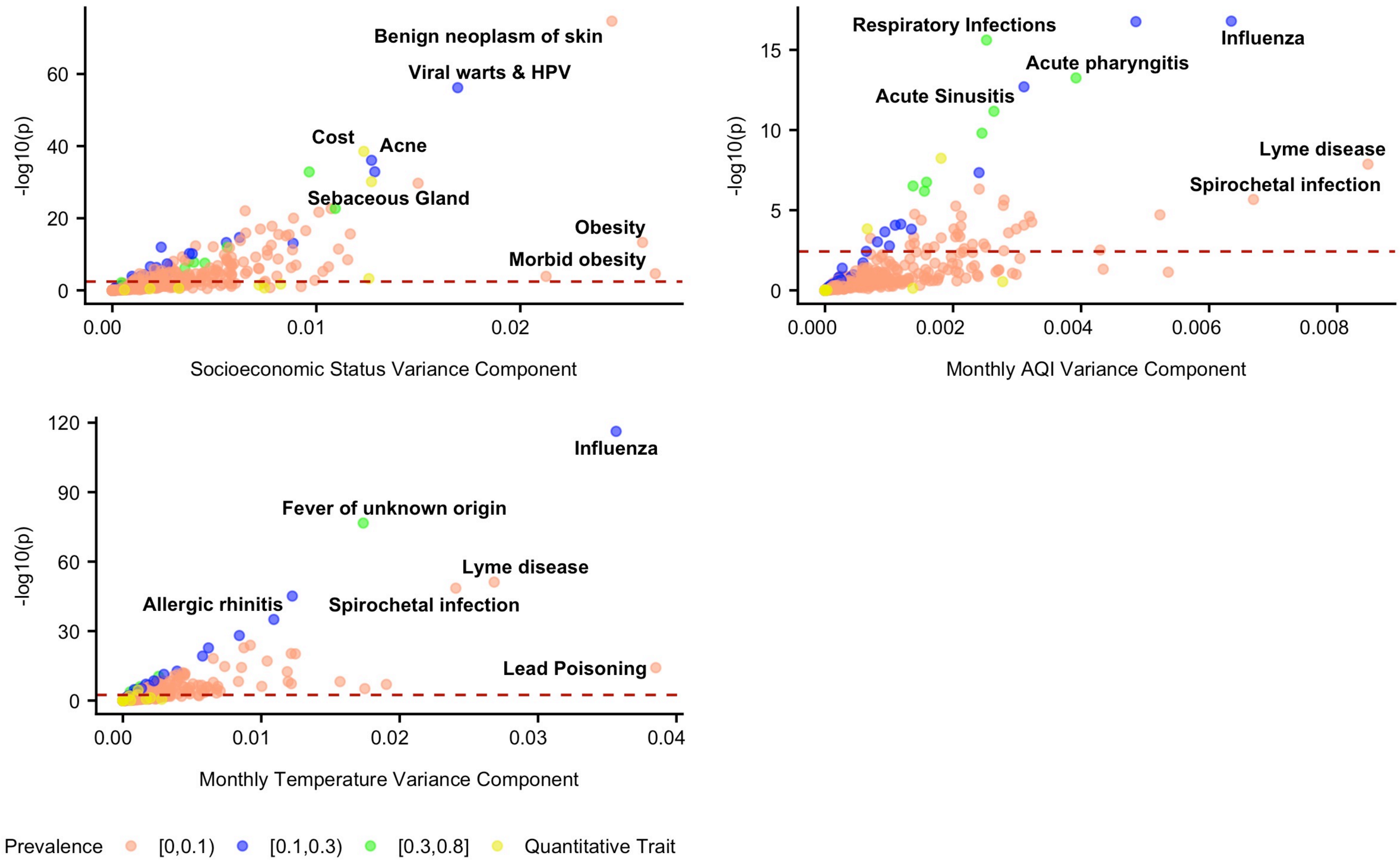


***h*² and *c*² estimates for *560* phenotypes versus statistical significance :**
326/560 traits (>50%) have a heritable and 180/560 (32%) had a shared exposome component!



Lakhani et al., Nature Genetics 2019

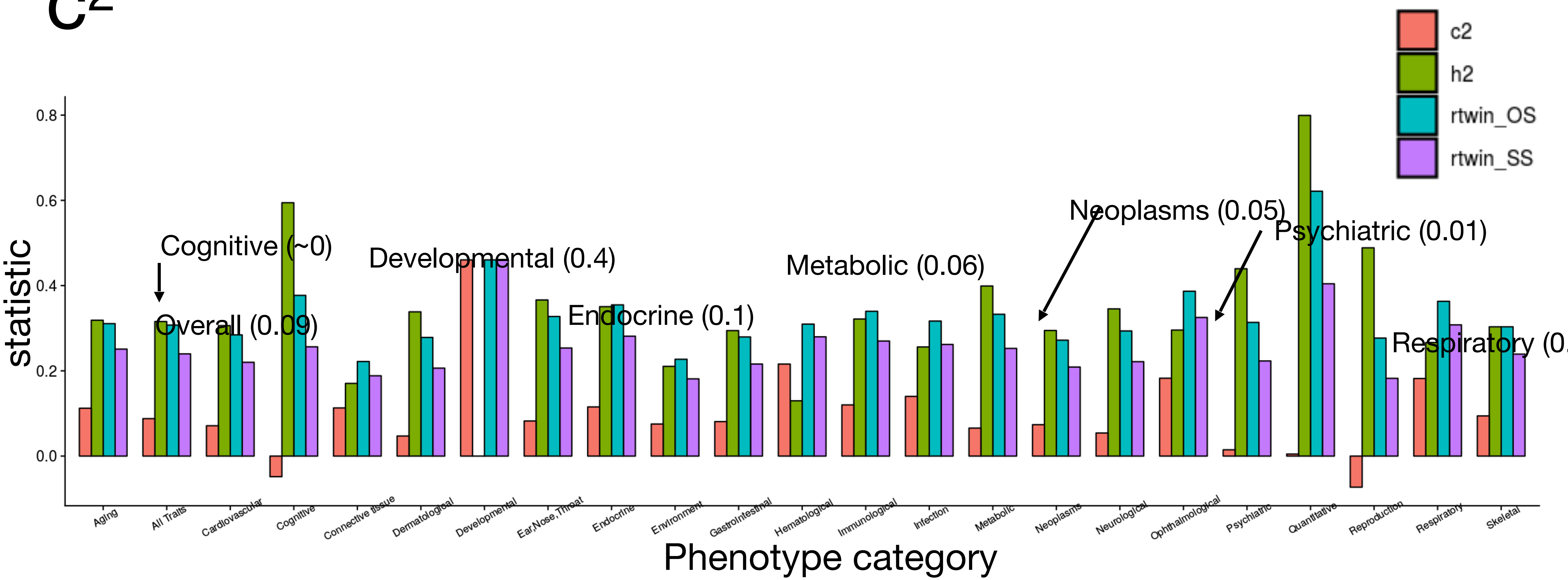
… but air pollution, climate, and SES played a modest role in total shared environment (c^2)



Lakhani et al., Nature Genetics 2019

Patient cohorts in the "real-world" :
*overall heritability (0.32) and **shared environment** (0.09)*

c^2



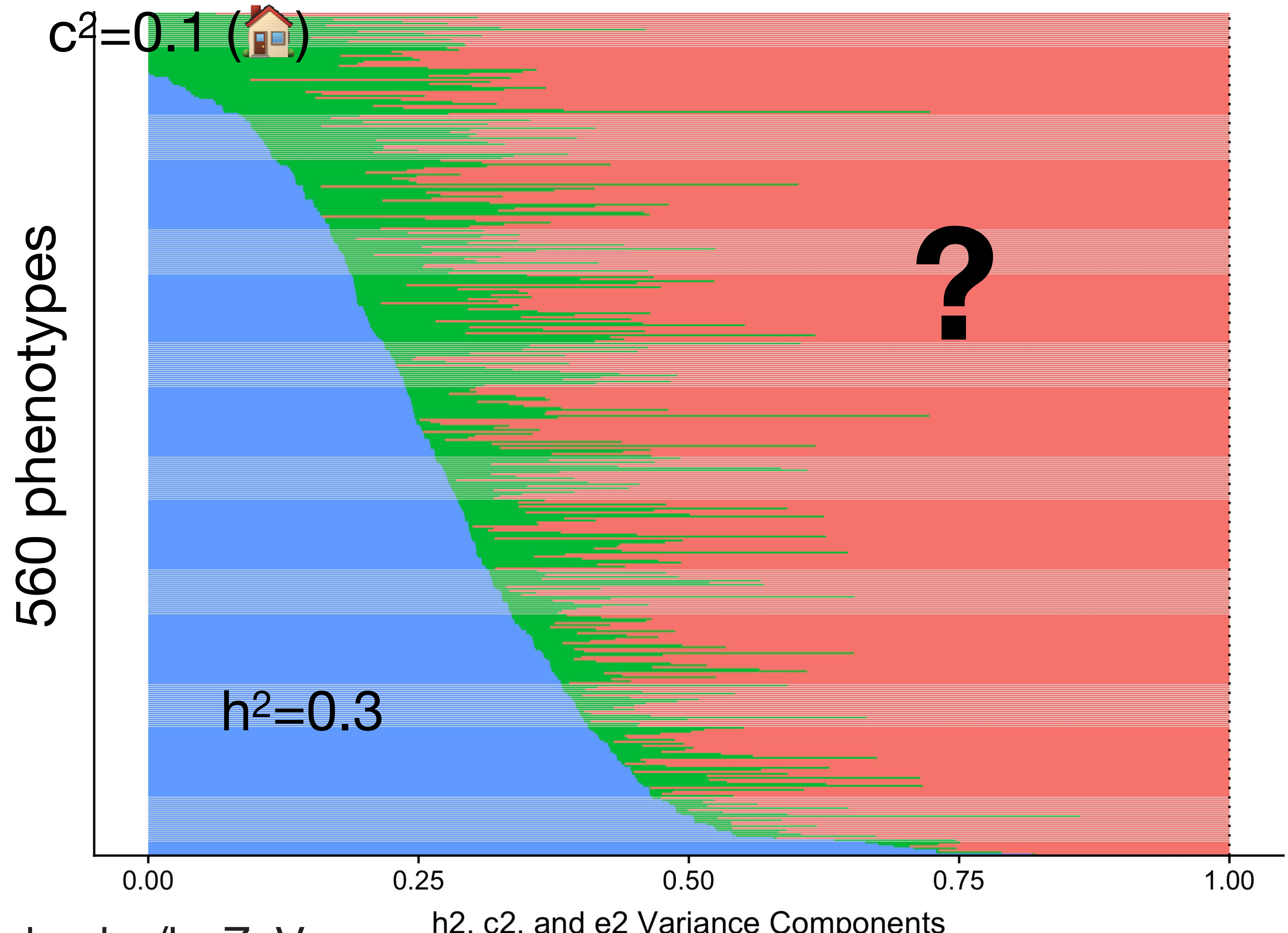
US-based, ages < 25

CaTCH: Claims analysis of Twin Correlation and Heritability

Lakhani et al., Nature Genetics 2019

http://apps.chiragjpgroup.org/catch/

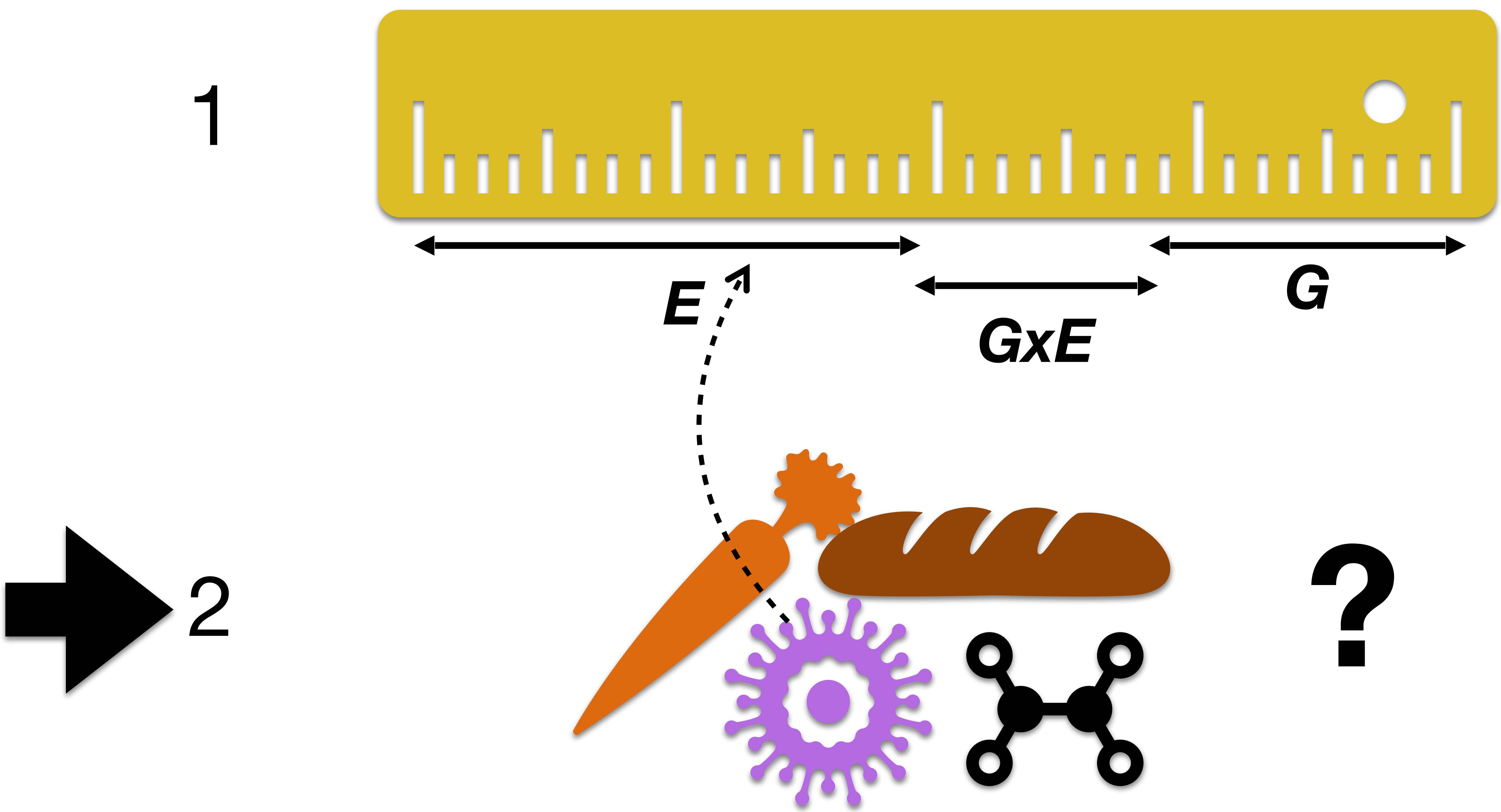
56K twins and 700K siblings in a massive health insurance cohort point to complex and elusive variation in 560 phenotypes



https://rdcu.be/boZeV

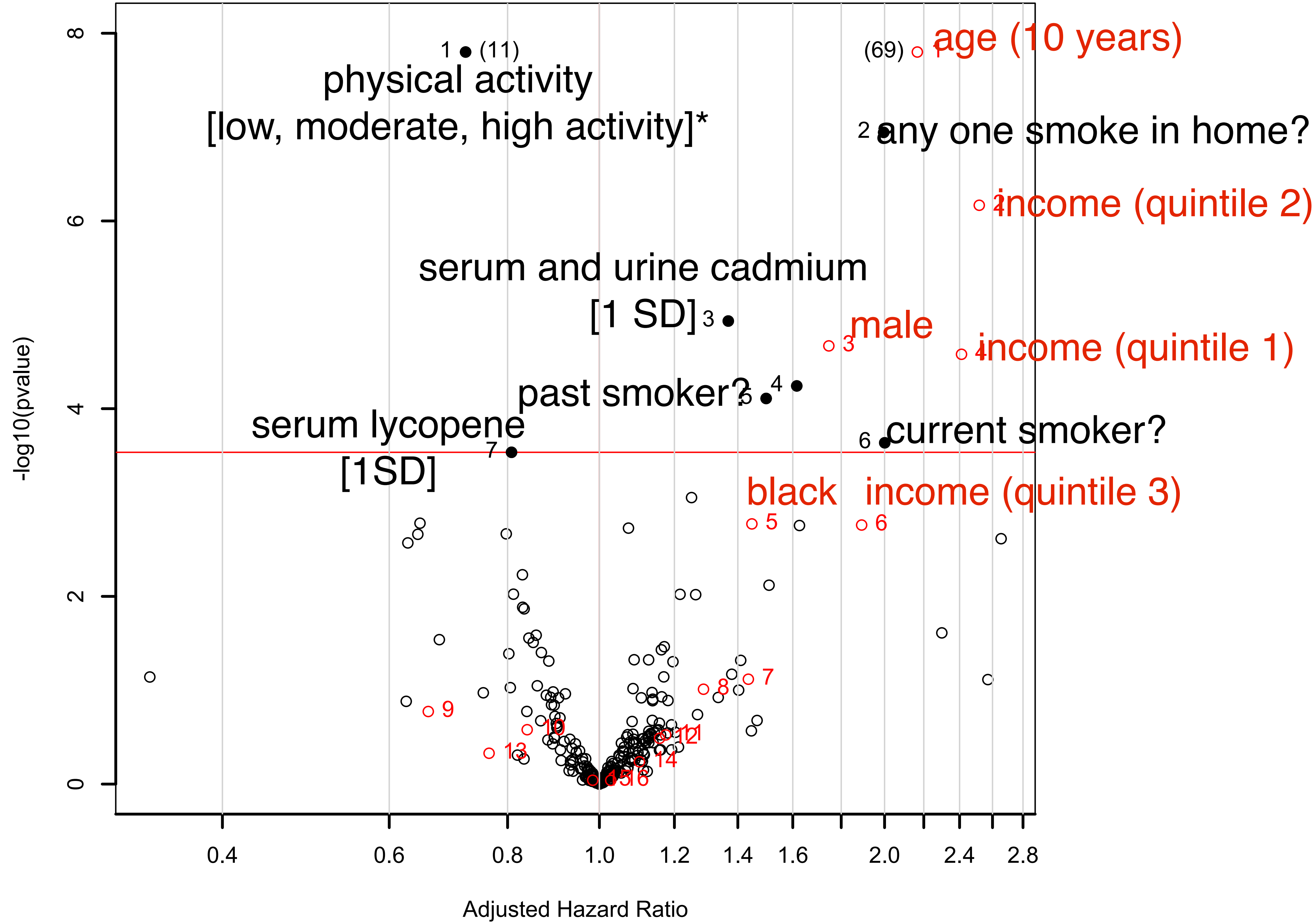
http://apps.chiragjpgroup.org/catch/

Exposomic research for precision medicine:
Measuring ***variance explained*** and accelerating ***discovery***



Exposome-wide association studies:
Data-driven discovery
Address multiplicity (False Discovery Rate)
Contextualize over multiple domains
(e.g., is *X* >> *Y*)?

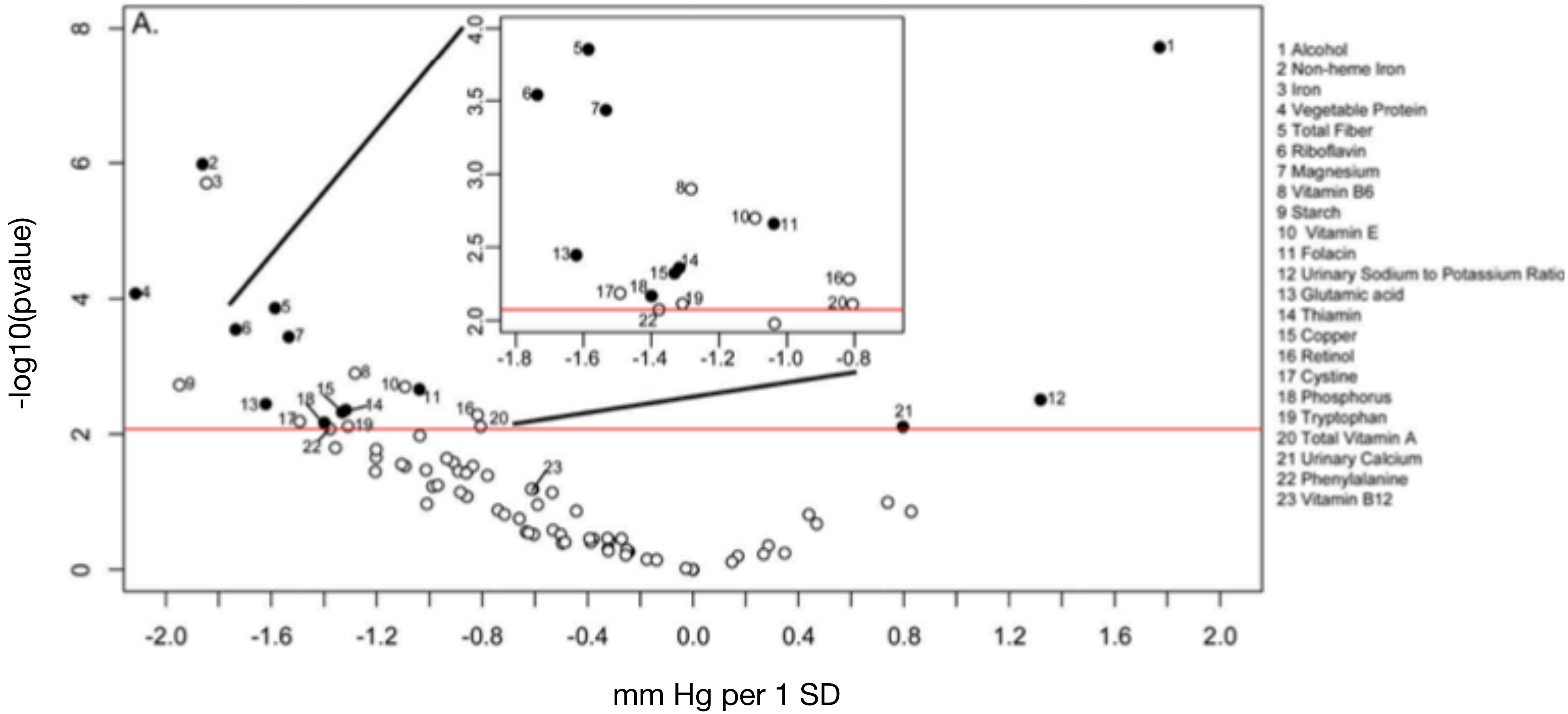
EWAS (re)-identifies factors associated with ***all-cause mortality***:
Volcano plot of 200 associations, incl. serum nutrients



Multivariate cox (age, sex, income, education, race/ethnicity, occupation [in red])
*derived from METs per activity and categorized by Health.gov guidelines

IJE 2012

Querying 85 nutrients across INTERMAP and NHANES in blood pressure: multiple factors emerge, but modest size



Range of 0.9-2 mm Hg per 1 SD change of a nutrient levels

Circulation 2012

Modest association sizes/variance for 500 *E* factors illuminates the need for consideration of exposome in aggregate

An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus

Chirag J. Patel[1,2,3], Jayanta Bhattacharya[4], Atul J. Butte[1,2,3]*

PLOS ONE 2010
ORs: 1.4-4 per 1 SD
4 identified (2%)

262 *E*

Systematic correlation of environmental exposure and physiological and self-reported behaviour factors with leukocyte telomere length

Chirag J. Patel,* Arjun K. Manrai, Erik Corona, and Isaac S. Kohane

461 *E*

IJE 2017
0-0.03 years (per 1SD)
21 identified (5%)

Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels

Chirag J Patel,[1,2] Mark R Cullen,[3] John PA Ioannidis[4,5,6] and Atul J Butte[1,2]*

IJE 2012
HDL-C: 1-3 mg/dL per 1 SD
LDL-C: 8-10 mg/dL per 1 SD
Triglycerides: 13-55 mg/dL per 1SD
~ 29 identified (10%)

188 *E*

Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States National Health and Nutrition Examination Survey

Chirag J Patel,[1] David H Rehkopf,[2] John T Leppert,[3] Walter M Bortz,[4] Mark R Cullen,[2] Glenn M Chertow[4] and John PA Ioannidis[1]*

249 *E*

IJE 2013
HRs: 0.7-2.8 (per 1SD)
7 identified (3%)

A systematic comprehensive longitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease

Soodabeh Milanlouei[1,5], Giulia Menichetti[1,5], Yanping Li[2], Joseph Loscalzo[3], Walter C. Willett[2,3] & Albert-László Barabási[1,3,4]

374 *E*

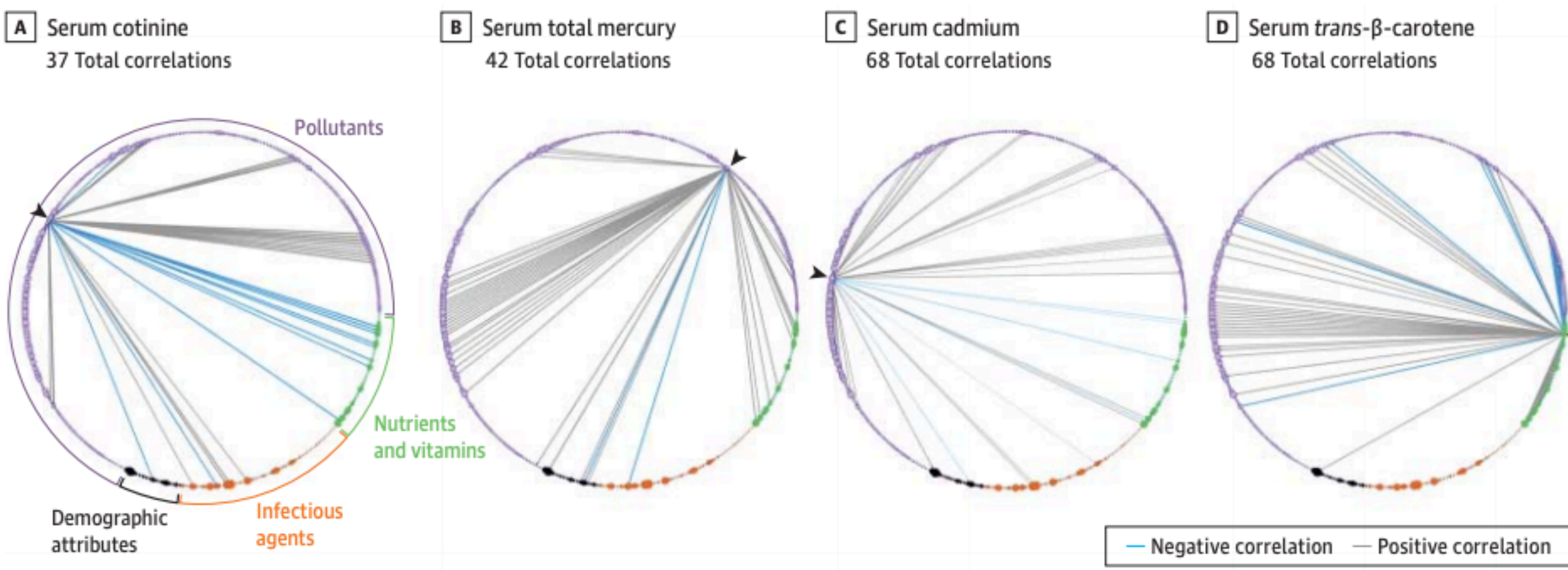
Nature Communications 2020
HRs: 0.9-1.3 per 1 SD
38 identified (10%)

Exposome-wide association study of semen quality: Systematic discovery of endocrine disrupting chemical biomarkers in fertility require large sample sizes

Ming Kei Chung[a], Germaine M. Buck Louis[b,c], Kurunthachalam Kannan[d], Chirag J. Patel[a,*]

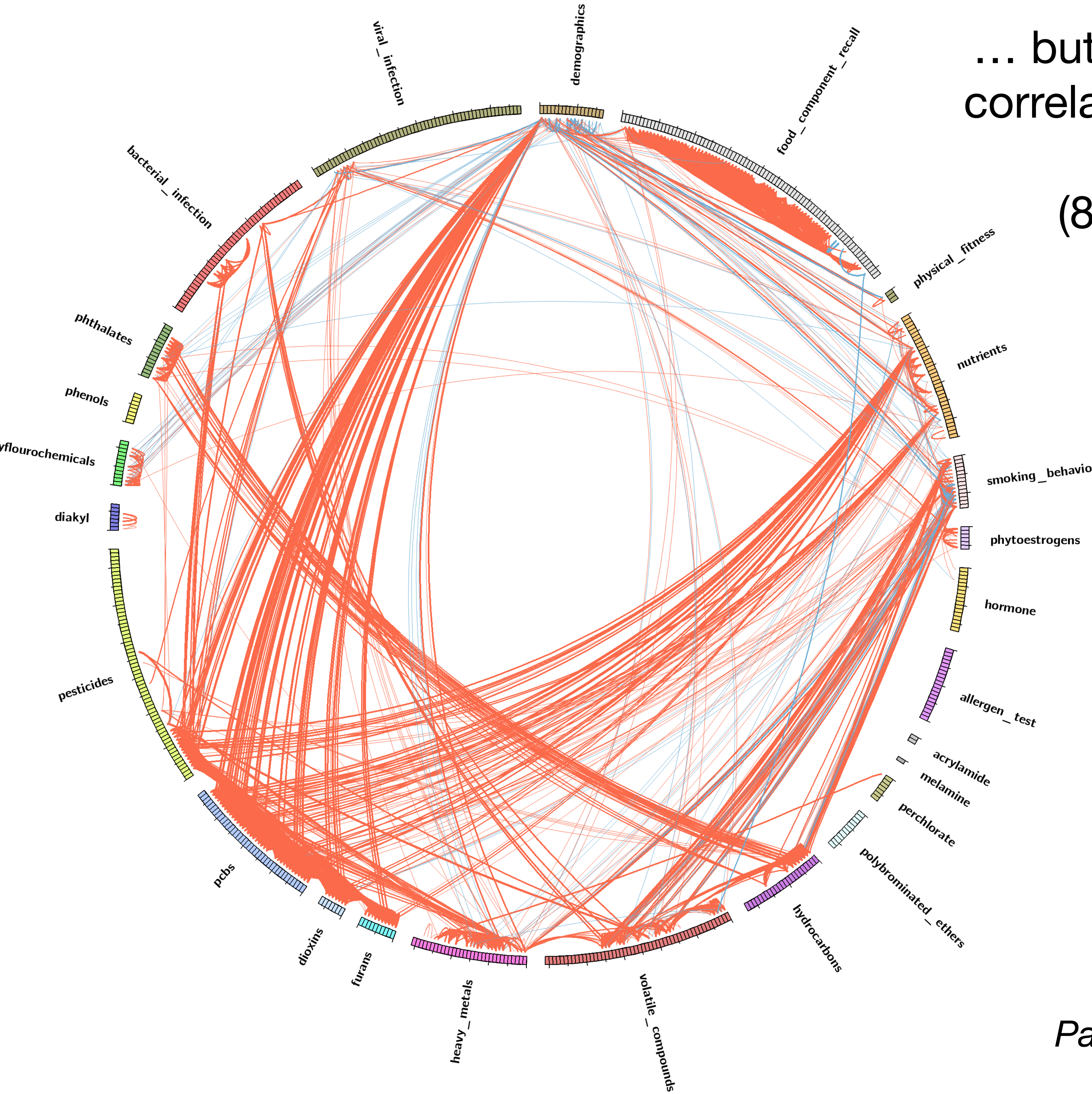
Env Int 2018
128 *E*
0 identified (0%)

Correlation structure between *E* factors: Correlation “globes” for 4 factors is dense…



JAMA 2014

… but modest: spearman correlation of < 0.3 among 575 ***E*** factors (82k correlations)



Pac Symp Biocomput. 2015
JECH 2015

A *maze* of associations is one way to a ***fragmented*** literature and ***Vibration of Effects***

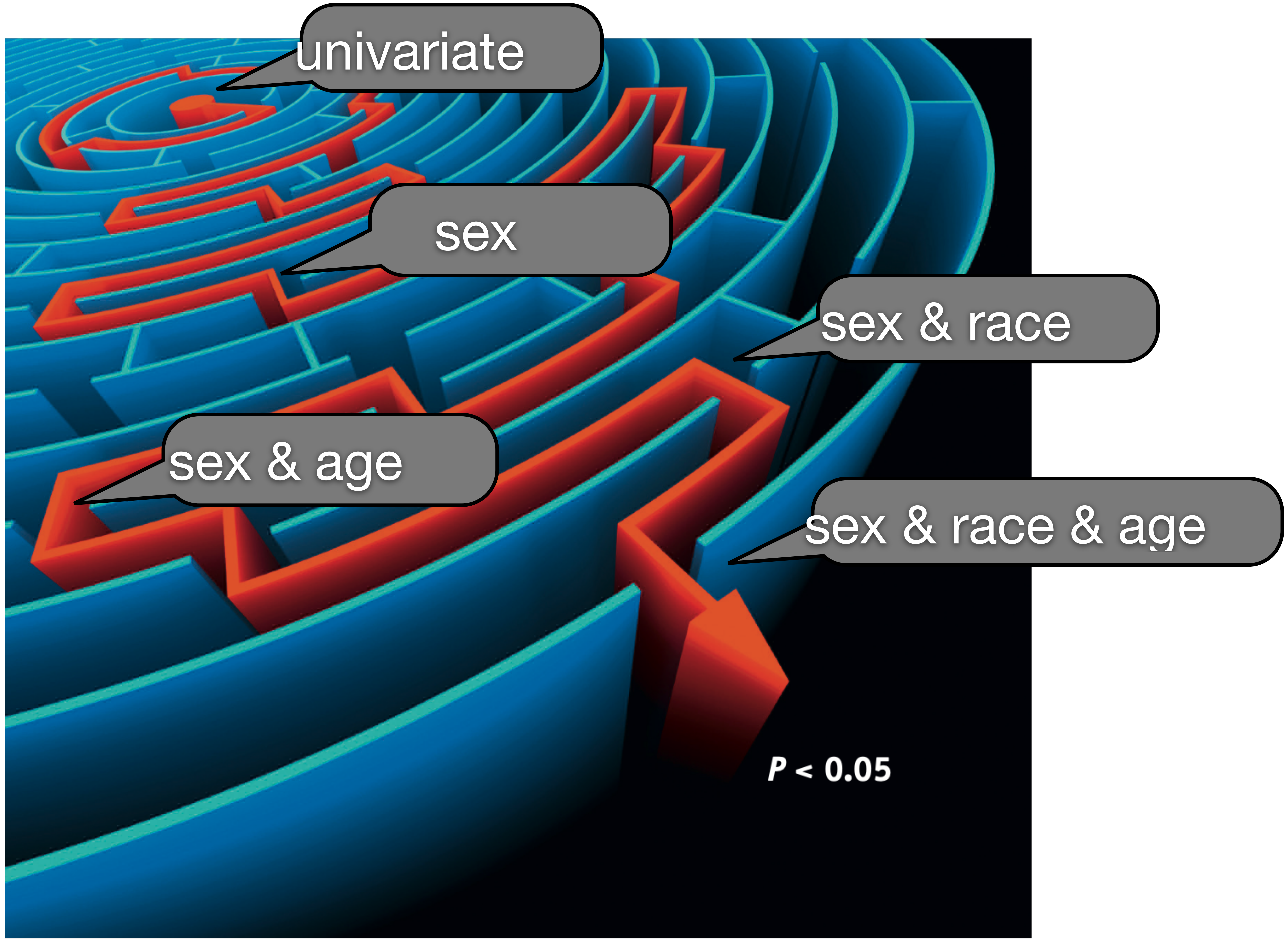


Figure 3. The path through a complex process can appear quite simple once the path is defined. Which terms are included in a multiple linear regression model? Each turn in a maze is analogous to including or not a specific term in the evolving linear model. By keeping an eye on the p-value on the term selected to be at issue, one can work towards a suitably small p-value. © ktsdesign – Fotolia

Young, 2011
JCE, 2015

See also: Gelman and Loken, 2013

Distribution of associations and p-values due to model choice: Estimating the ***Vibration of Effects (or Risk)***

A

Data Source
NHANES 1999-2004
417 variables of interest
time to death
N≧1000 (≧100 deaths)

B

Variable of Interest
e.g., 1 SD of log(serum Vitamin D)

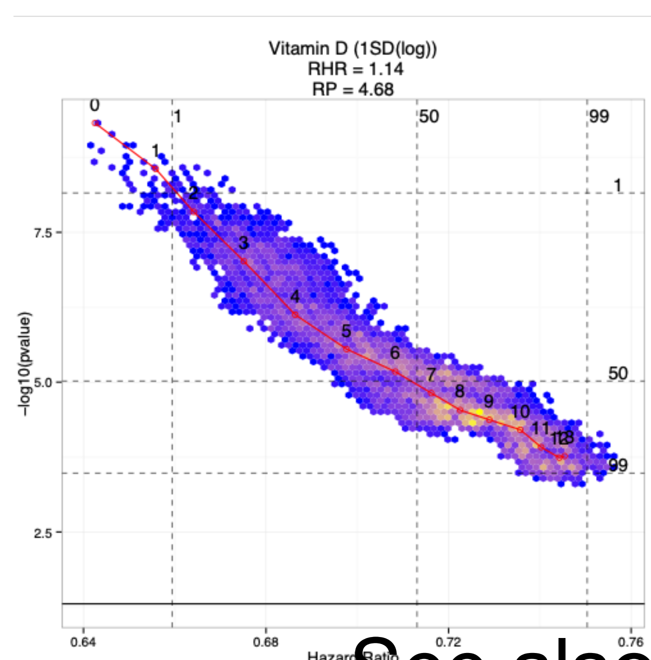
C

Adjusting Variable Set
n=13

SES [3rd tertile]
education [>HS]
race [white]
body mass index [normal]
total cholesterol
any heart disease
family heart disease
any hypertension
any diabetes
any cancer
current/past smoker [no smoking]
drink 5/day
physical activity

D

All-subsets Cox regression
$2^{13} + 1 = 8{,}193$ models

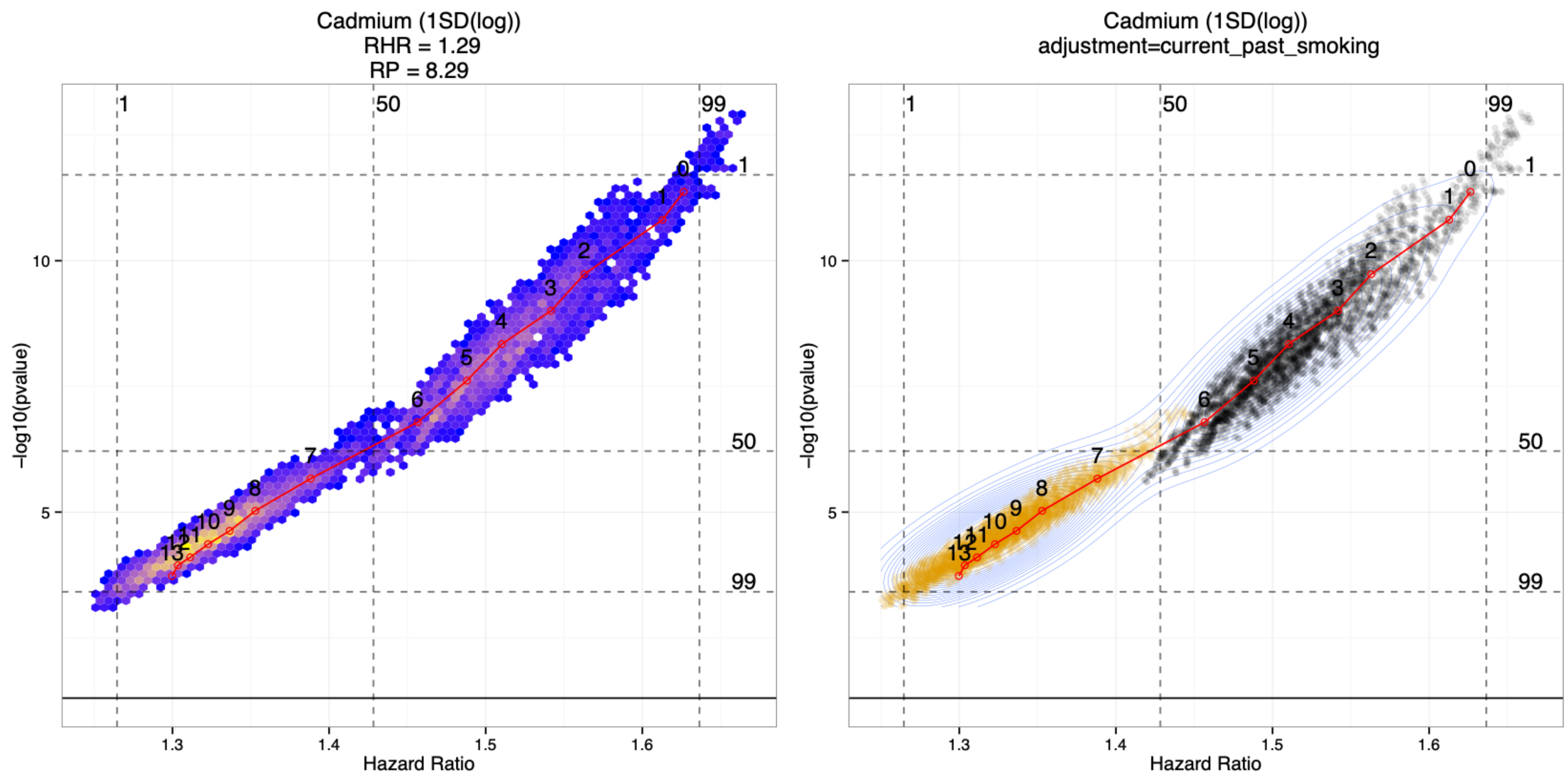


JCE, 2015

See also: Gelman and Loken, 2013
Young, 2011

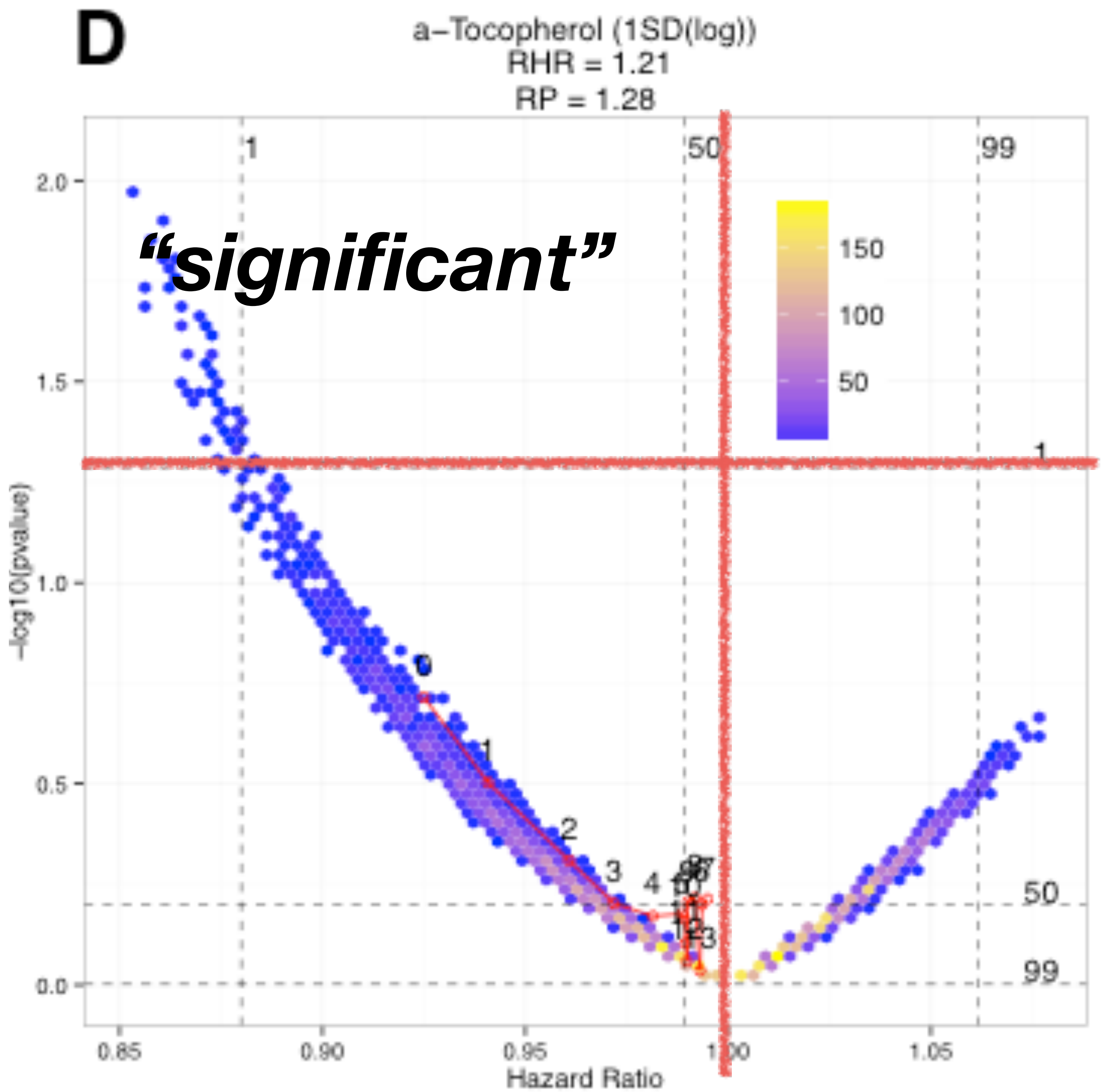
http://bit.ly/effectvibration

The ***Vibration of Effects:*** shifts in the effect size distribution due to select adjustments (e.g., adjusting ***cadmium levels) with smoking status***)



JCE, 2015

The ***Vibration of Effects:*** beware of the *Janus* effect
(both ***risk*** *and* ***protection****?!)*



"protection" *"risk"*

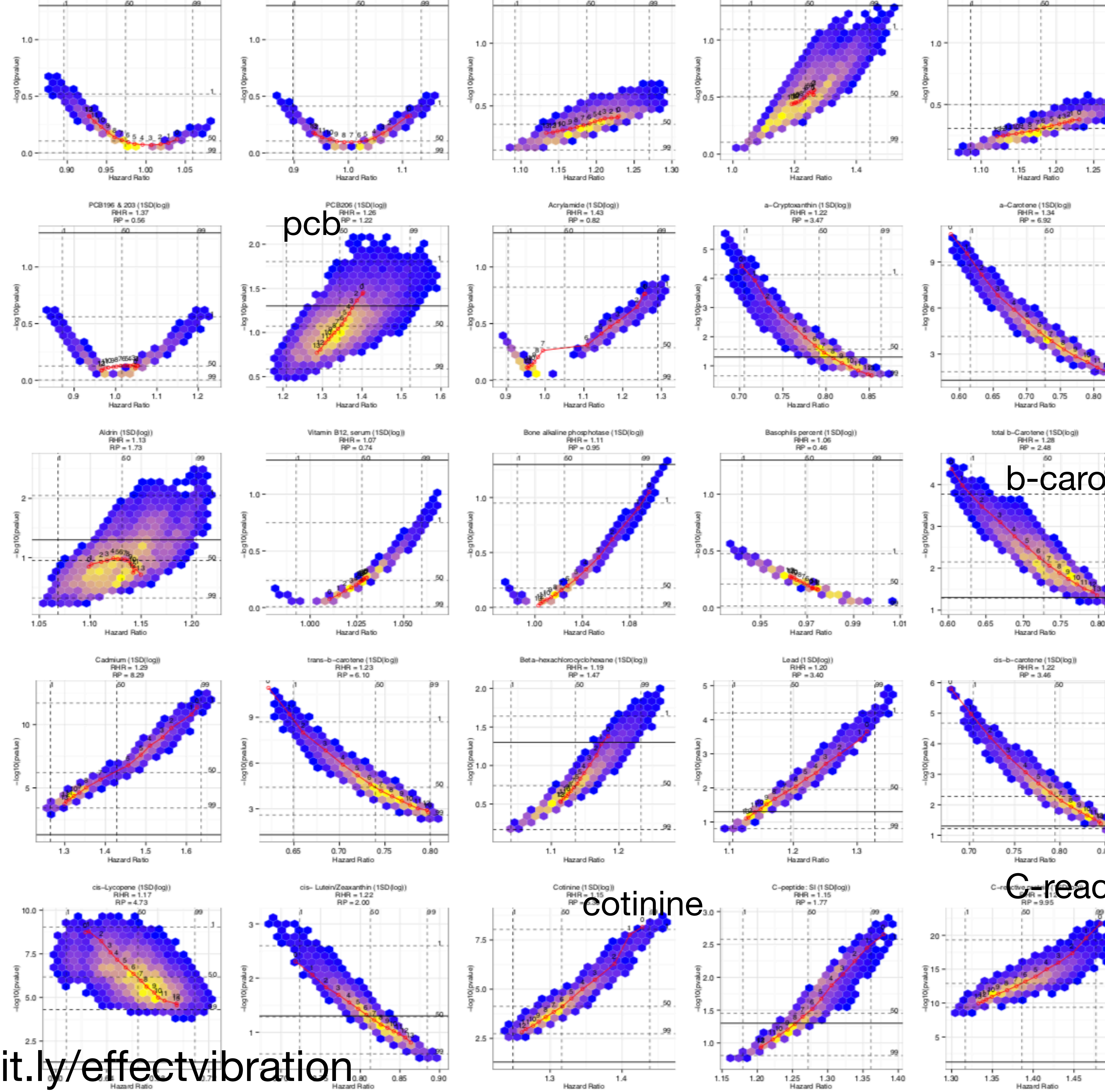
Brittanica.com

Janus (two-faced) risk profile

Risk and ***significance*** depends on modeling scenario -
Where variation is smallest may be the most robust

http://bit.ly/effectvibration

JCE, 2015



http://bit.ly/effectvibration

The ***polygenic risk score*** (PRS) has emerged as a way of summarizing ***additive genetic risk*** for ***time-invariant screening***

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4*]

Khera et al, Nature Genetics 2018

See also:
Choi et al, Nat Protoc 2020
Vilhjálmsson et al, AJHG 2015

Viewpoint

April 3, 2020

Addressing Social Determinants of Health

Time for a Polysocial Risk Score

Jose F. Figueroa, MD, MPH[1,2]; Austin B. Frakt, PhD[1,3,4]; Ashish K. Jha, MD, MPH[1,5]

☐ Author Affiliations | Article Information

JAMA. 2020;323(16):1553-1554. doi:10.1001/jama.2020.2436

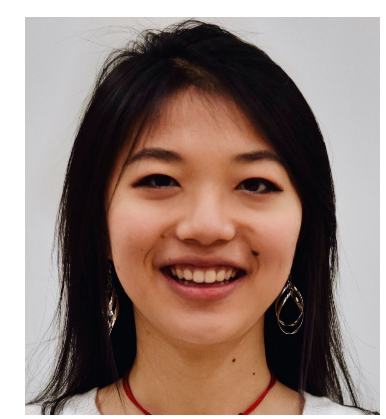
Attempts to link individual social factors to outcomes are similar to assuming disorders such as CAD are driven by a mendelian genetic disorder. Poorly controlled diabetes may be made worse by low-quality and insecure housing, but also could be affected by food insecurity, lack of social networks, inadequate health literacy, and other social factors that also are interlinked. Efforts to quantify the contribution of any single factor on health outcomes are often not helpful.

Social determinants have become a catch-all for any factor outside the traditional health care system that is correlated with health. While a variety of social factors influence health, these factors do so in complex and interrelated ways. Viewing these factors in isolation leads to interventions applicable to only a minority of patients (ie, providing housing for the homeless) and, although important, are limited. Polysocial risk scores offer a possible approach to assess the influence of social factors on disease and health outcomes more broadly, and may have the potential to help target health care interventions, programs, and resources more effectively across wider populations.

Figueroa et al. *JAMA* 2020

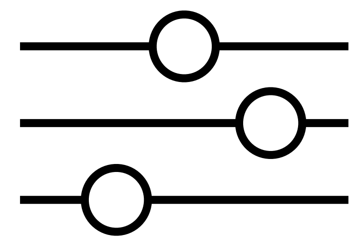
Building a ***P***oly-e**X**posure Risk **S**core (***PXS***)

N=111
Accommodations
Air pollution
Alcohol
Diet
Early life factors
Education
Employment
Income
Lifestyle/Exercise
Sociodemographics
Sleep
Smoking
Sound pollution



Yixuan He
Diabetes Care 2021

Building a ***P**oly-e**X**posure Risk **S**core (**PXS**)*



N=111
Accommodations
Air pollution
Alcohol
Diet
Early life factors
Education
Employment
Income
Lifestyle/Exercise
Sociodemographics
Sleep
Smoking
Sound pollution

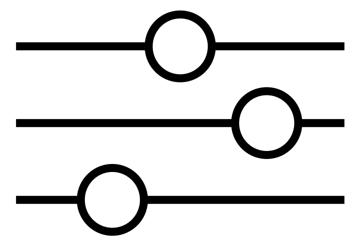
Filter & Select
XWAS
Lasso
P value thresholds

Diabetes Care 2021

Building a ***P***oly-e**X**posure Risk **S**core (***PXS***)

N=111

Accommodations
Air pollution
Alcohol
Diet
Early life factors
Education
Employment
Income
Lifestyle/Exercise
Sociodemographics
Sleep
Smoking
Sound pollution



Filter & Select

XWAS
Lasso
P value thresholds

N=12

Alcohol intake
Comparative body size at age 10
Major dietary changes in past five years
Household income
Insomnia
Snoring
Milk type used (skim, whole, etc.)
Dietary restriction (eggs, diary, wheat, etc)
Spread type used (butter, etc)
Tea intake per day
Own or rent accommodations
Past tobacco usage

PXS: C statistics of 0.762 (0.749, 0.775)
PGS: C statistic of 0.709 (0.696, 0.722)
PXS may have utility in reclassification improvement of T2D beyond classical risk factors (BP, glucose, A1C, lipids)

A

	CRS+PGS Model		
CRS Model	# Participants	Continuous NRI	Categorical NRI
Cases	1281	0.152 (0.115 to 0.191)	0.065 (0.021 to 0.118)
Noncases	67018	0.073 (0.055 to 0.092)	-0.005 (-0.009 to -0.002)
Full population	68299	0.225 (0.174 to 0.280)	0.060 (0.020 to 0.109)

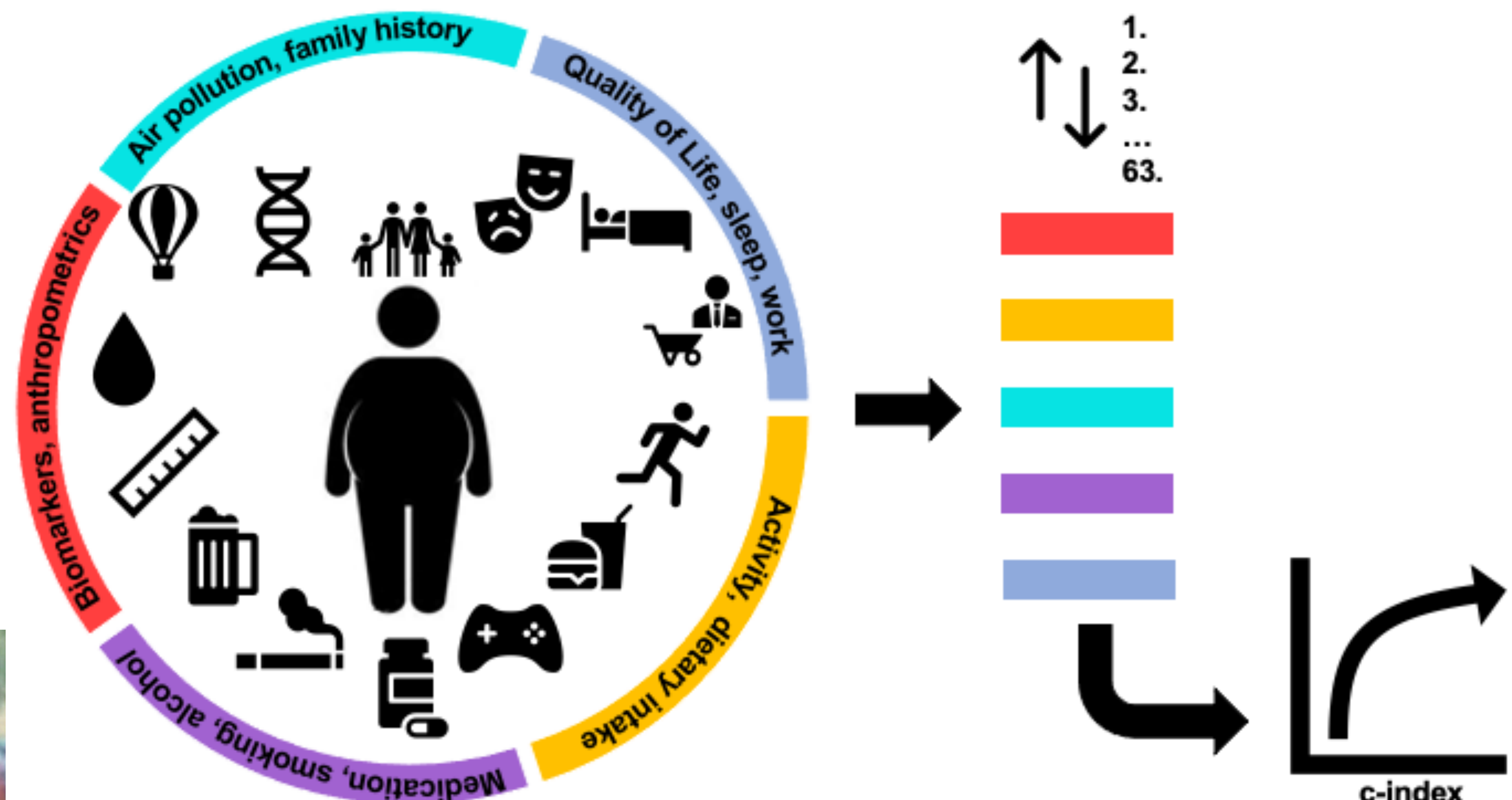
B

	CRS+PXS Model		
CRS Model	# Participants	Continuous NRI	Categorical NRI
Cases	1281	0.301 (0.259 to 0.336)	0.091 (0.033 to 0.154)
Noncases	67018	0.169 (0.144 to 0.193)	-0.005 (-0.011 to -0.001)
Full population	68299	0.470 (0.406 to 0.523)	0.085 (0.032 to 0.144)

PXS+classical risk factors: C statistic of 0.850
Noble et al.: AUC 0.6-0.9 (BMJ, 2011)
Meigs et al.: C-index 0.9 (NEJM, 2008)

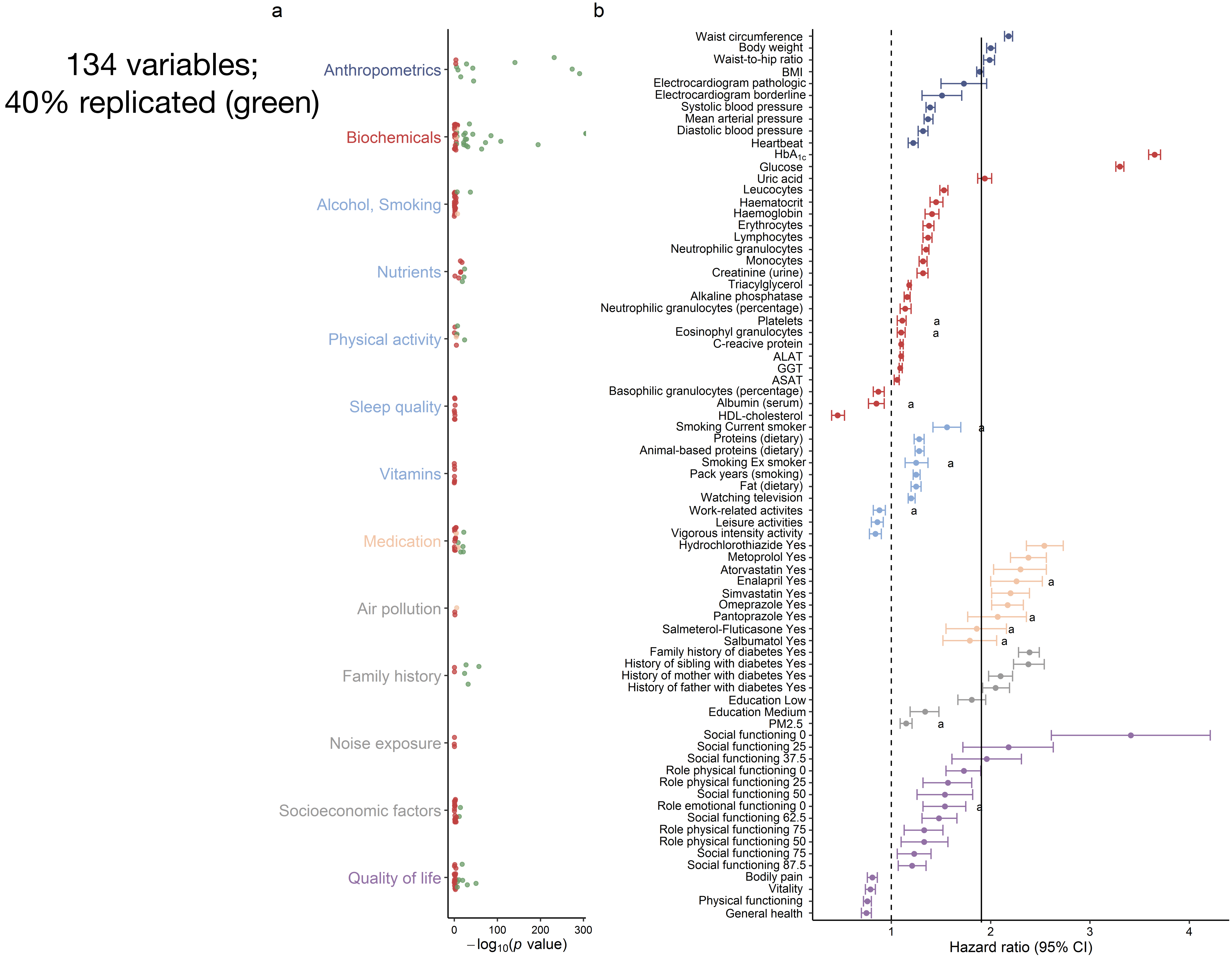
Diabetes Care 2021

How ***interchangeable*** are E factors in diagnosed T2D risk? Contextualizing 134 variables in *LIFELINES* (Netherlands, n=95k)



Tom van der Meer

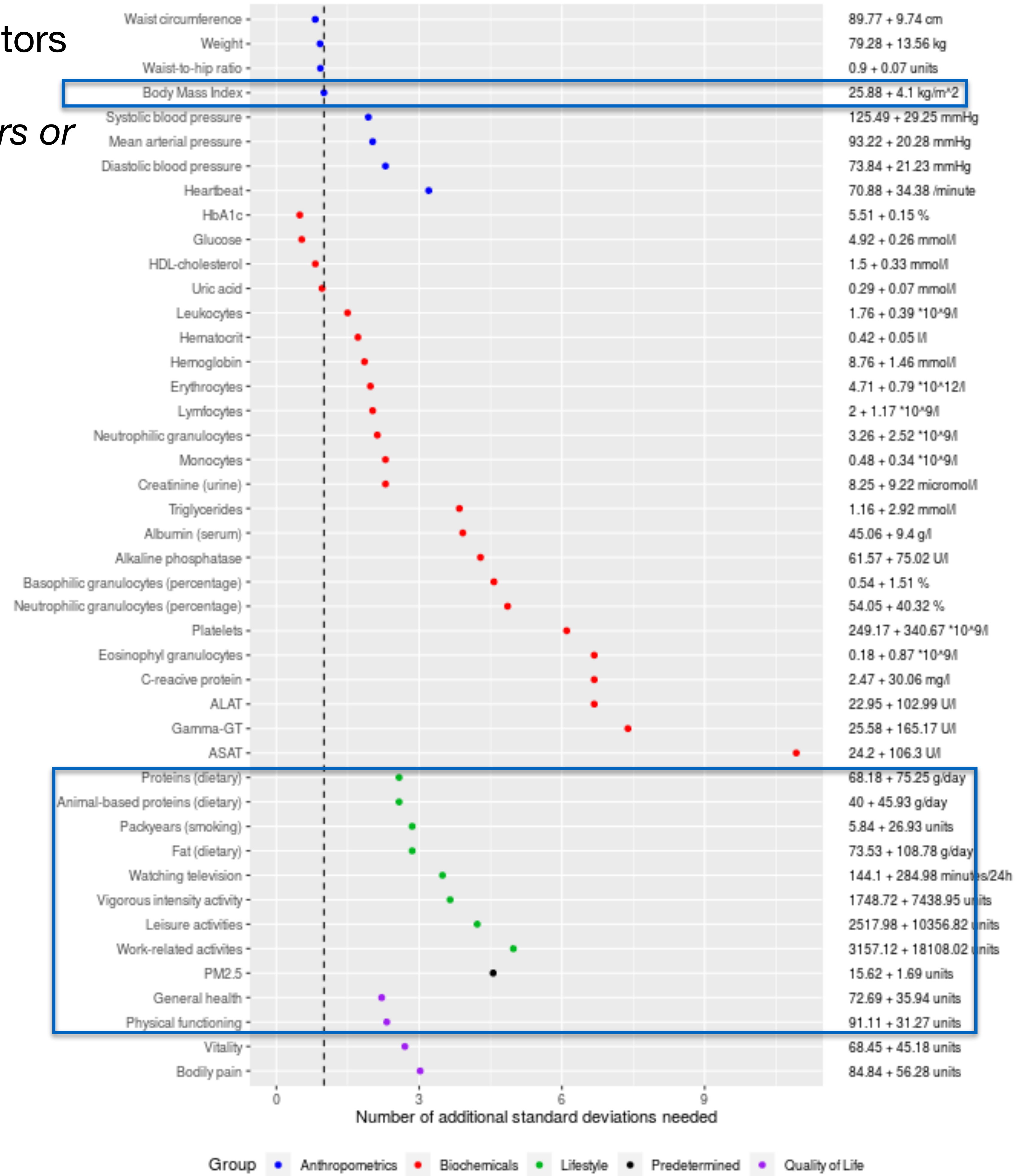
Diabetologia 2021



a
b
134 variables;
40% replicated (green)
Anthropometrics
Biochemicals
Alcohol, Smoking
Nutrients
Physical activity
Sleep quality
Vitamins
Medication
Air pollution
Family history
Noise exposure
Socioeconomic factors
Quality of life
0
100
200
300
$-\log_{10}(p \text{ value})$
Waist circumference
Body weight
Waist-to-hip ratio
BMI
Electrocardiogram pathologic
Electrocardiogram borderline
Systolic blood pressure
Mean arterial pressure
Diastolic blood pressure
Heartbeat
HbA_{1c}
Glucose
Uric acid
Leucocytes
Haematocrit
Haemoglobin
Erythrocytes
Lymphocytes
Neutrophilic granulocytes
Monocytes
Creatinine (urine)
Triacylglycerol
Alkaline phosphatase
Neutrophilic granulocytes (percentage)
Platelets
Eosinophyl granulocytes
C-reacive protein
ALAT
GGT
ASAT
Basophilic granulocytes (percentage)
Albumin (serum)
HDL-cholesterol
Smoking Current smoker
Proteins (dietary)
Animal-based proteins (dietary)
Smoking Ex smoker
Pack years (smoking)
Fat (dietary)
Watching television
Work-related activites
Leisure activities
Vigorous intensity activity
Hydrochlorothiazide Yes
Metoprolol Yes
Atorvastatin Yes
Enalapril Yes
Simvastatin Yes
Omeprazole Yes
Pantoprazole Yes
Salmeterol-Fluticasone Yes
Salbumatol Yes
Family history of diabetes Yes
History of sibling with diabetes Yes
History of mother with diabetes Yes
History of father with diabetes Yes
Education Low
Education Medium
PM2.5
Social functioning 0
Social functioning 25
Social functioning 37.5
Role physical functioning 0
Role physical functioning 25
Social functioning 50
Role emotional functioning 0
Social functioning 62.5
Role physical functioning 75
Role physical functioning 50
Social functioning 75
Social functioning 87.5
Bodily pain
Vitality
Physical functioning
General health
1
2
3
4
Hazard ratio (95% CI)

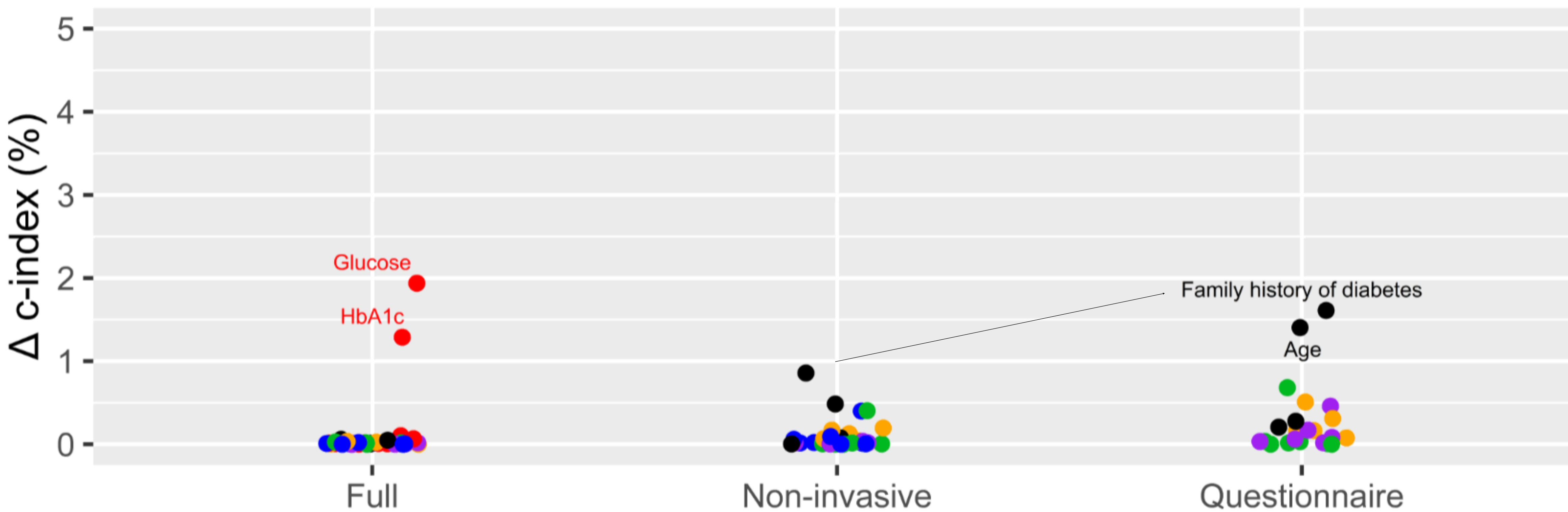
Comparing potential risk factors against BMI

~= 3 SDs of modifiable factors or higher



Diabetologia 2021

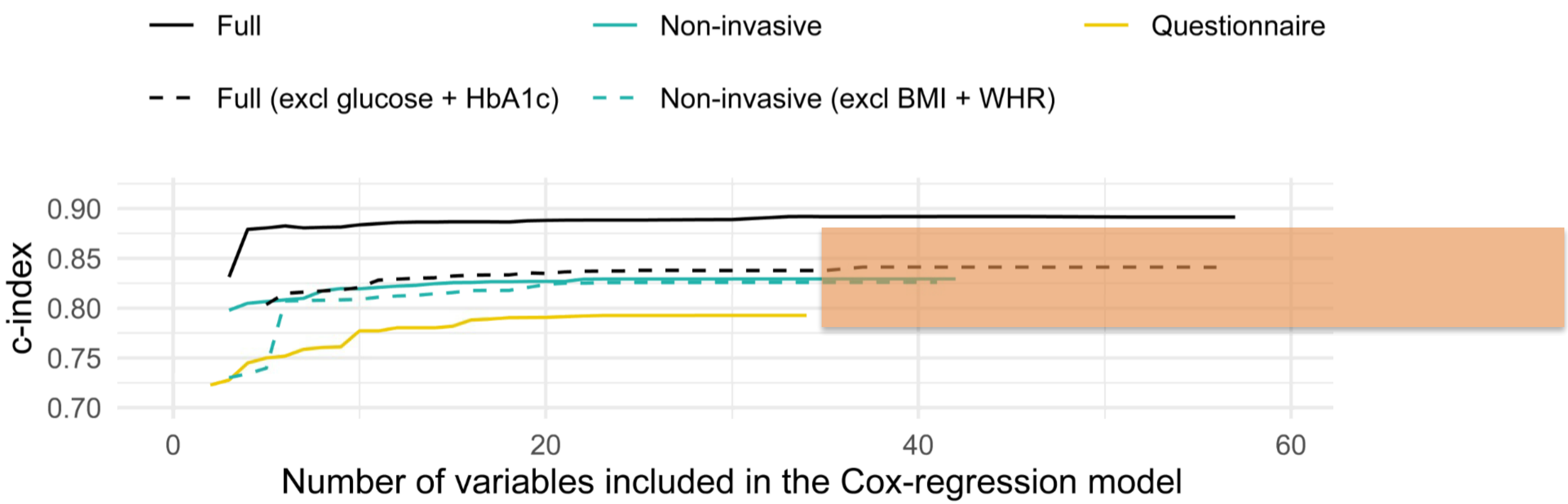
Equal feature importance of most *E* and *P* variables point to their interchangeability



Glucose, A1C%, family history contribute uniquely
All others variables are interchangeable

Diabetologia 2021

Risk prediction of T2D satiates after the inclusion of a limited number of *E* and *P* variables
(risk is dominated by a few key players and plateaus quickly)



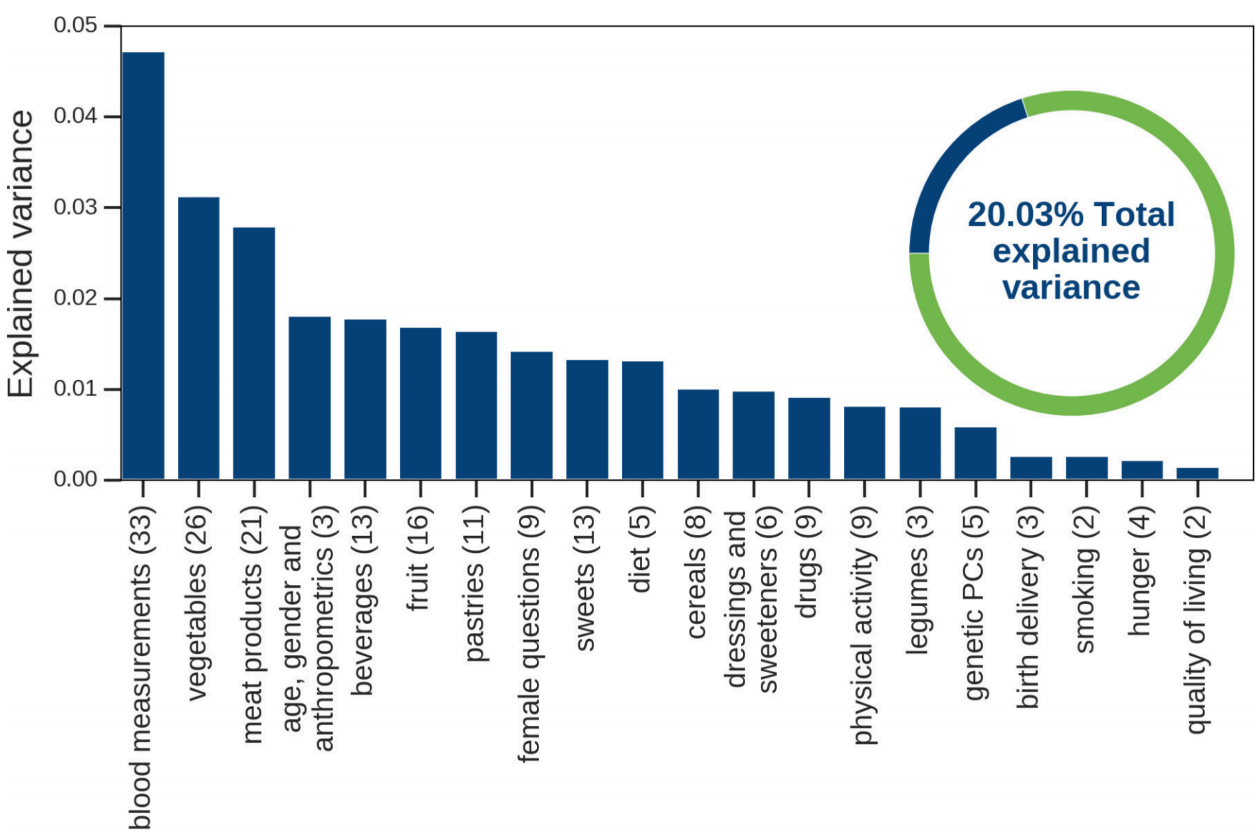
- Full and non-invasive models satiate after the inclusion of 4 variables
- Removing key non-invasive variables leads to larger models with similar prediction

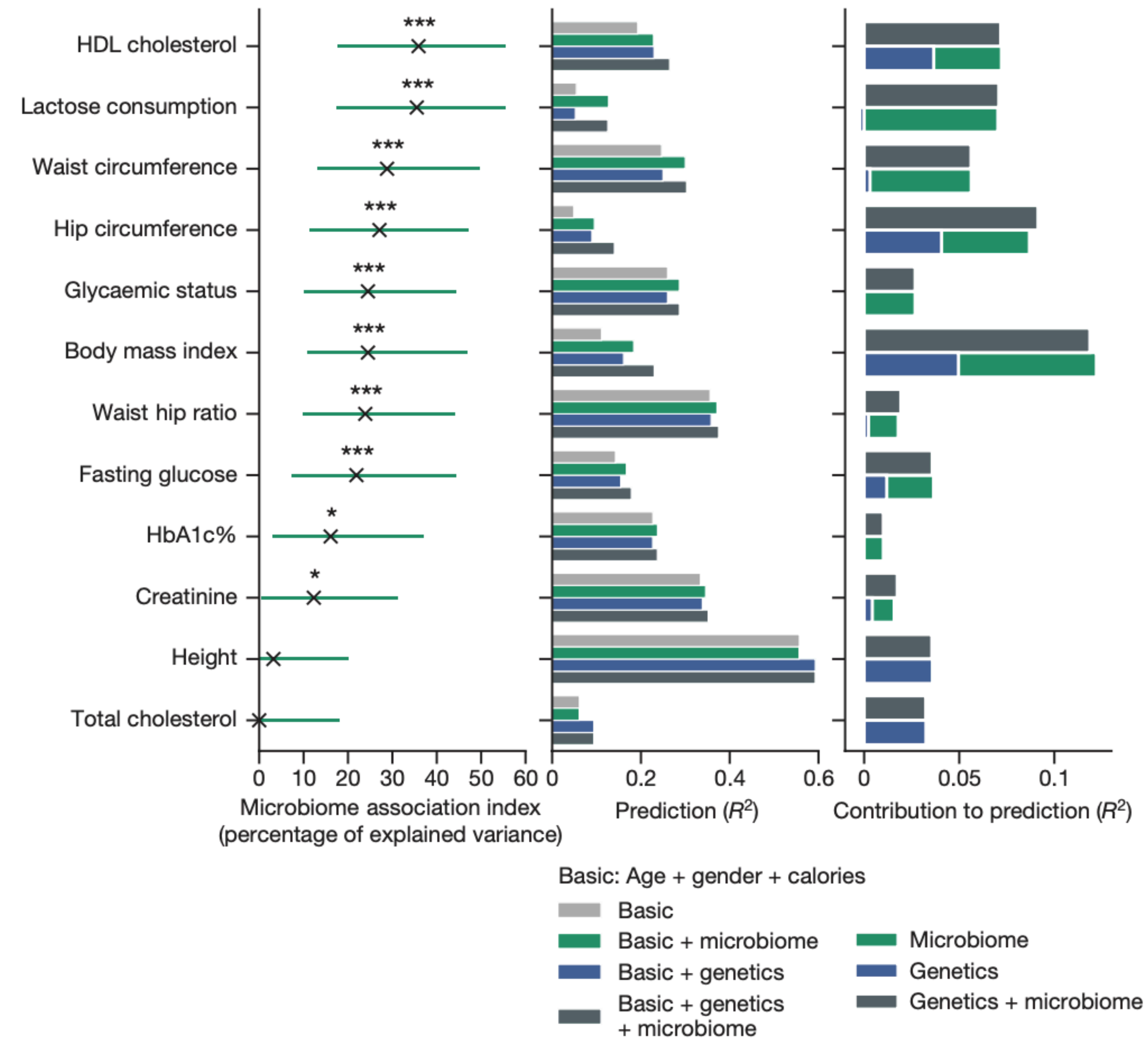
Diabetologia 2021

Environment dominates over host genetics in shaping human gut microbiota

Daphna Rothschild[1,2]*, Omer Weissbrod[1,2]*, Elad Barkan[1,2]*, Alexander Kurilshikov[3], Tal Korem[1,2], David Zeevi[1,2], Paul I. Costea[1,2], Anastasia Godneva[1,2], Iris N. Kalka[1,2], Noam Bar[1,2], Smadar Shilo[1,2], Dar Lador[1,2], Arnau Vich Vila[3,4], Niv Zmora[5,6,7], Meirav Pevsner-Fischer[5], David Israeli[8], Noa Kosower[1,2], Gal Malka[1,2], Bat Chen Wolf[1,2], Tali Avnit-Sagi[1,2], Maya Lotan-Pompan[1,2], Adina Weinberger[1,2], Zamir Halpern[7,9], Shai Carmi[10], Jingyuan Fu[3,11], Cisca Wijmenga[3,12], Alexandra Zhernakova[3], Eran Elinav[5]§ & Eran Segal[1,2]§

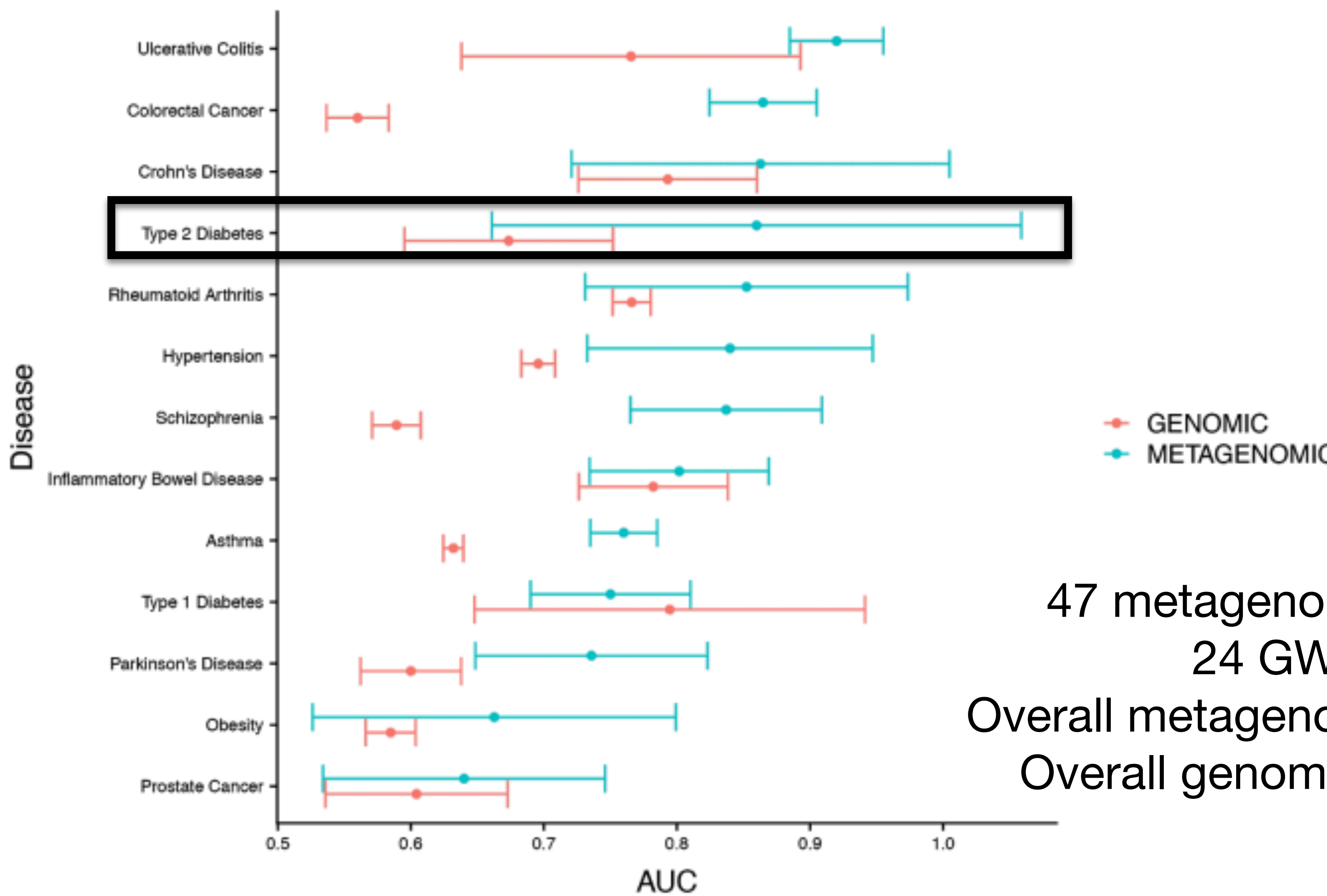
Nature 2018



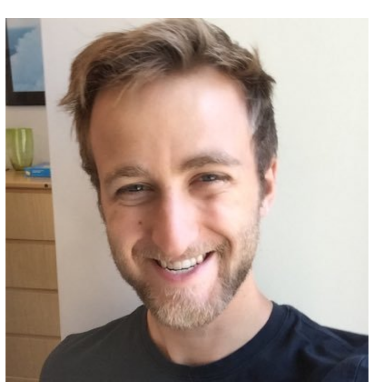


The predictive power of the microbiome exceeds that of genome-wide association studies in the discrimination of complex human disease

Braden T Tierney, Yixuan He, George M Church, Eran Segal, Aleksandar D Kostic, Chirag J Patel



47 metagenomic studies
24 GWASs
Overall metagenomic AUC: 0.79
Overall genomic AUC: 0.67

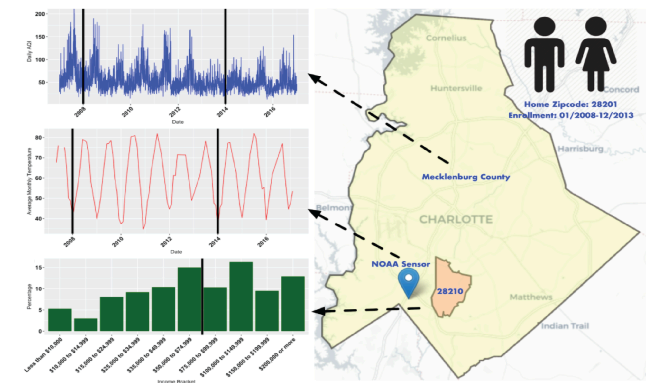


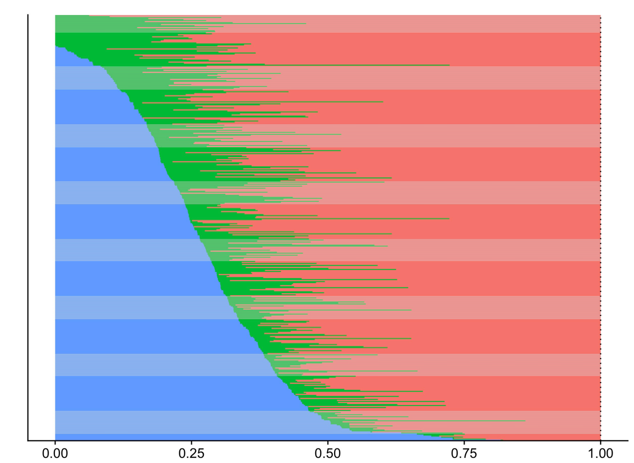
Braden Tierney

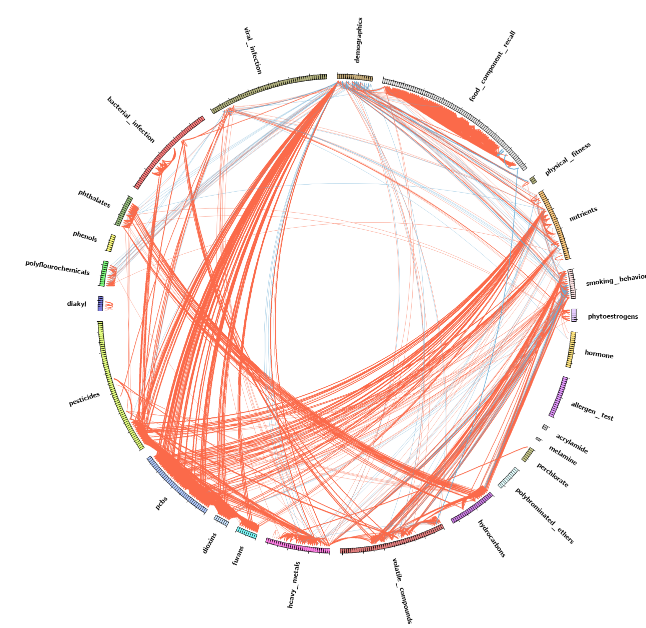
https://doi.org/10.1101/2019.12.31.891978

Architecture of phenotype-***exposome*** associations: New discovery vs. utility

- ***Promises***
  - ***Mapping of E onto P*** (discovery of new biology)
  - ***Measurement:*** *E at scale with low error through time via mass-spec and 'omics*
  - ***Decoupling shared*** and ***non-shared E and time course***
  - ***Polyexposure scores:*** Solving the ***architecture*** of mixtures of E-P
- ***Challenges***
  - ***"Real-world"/observational estimates are small***
  - ***Dense correlation:*** *everything correlated with everything else*
  - ***Uncertainty of models and black box predictors***
  - ***Multiplicity:*** *unclear how to prioritize needles among needles of factors for experiments*
- ***Outlook:***
  - Consideration of the massive expanse of E needed to describe variation
  - **Predictive advances with new big data approaches will be rare**: *last mile of AUC and R^2 as risk factors may be interchangable*







Acknowledgements

RagGroup
Chirag Lakhani
Yixuan He
Tom Van Der Meer
Braden Tierney
Jake Chung
Jason Tan
Kajal Claypool
Alan LeGoallec
Danielle Rasooly
Andrew Deonarine

Collaborators Mentioned
Arjun Manrai
John Ioannidis
Peter Visscher
Alexsander Kostic

HARVARD MEDICAL SCHOOL
BLAVATNIK INSTITUTE BIOMEDICAL INFORMATICS

Chirag J Patel
chirag@hms.harvard.edu
@chiragjp
www.chiragjpgroup.org