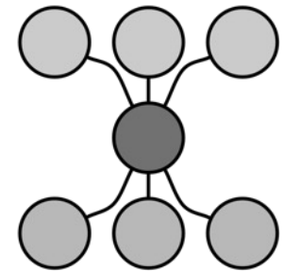


High-throughput epidemiology: Contrasting the genome and exposome at biobank scale

Chirag J Patel

Current Issues in Genomics and Precision Public Health
Oak Ridge Institute for Science and Education Training
Atlanta, Georgia
9/8/2023

HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics

chirag@hms.harvard.edu
@chiragjp
www.chiragjpgroup.org

Phenome *Genome* *Exposome*

P = G + E

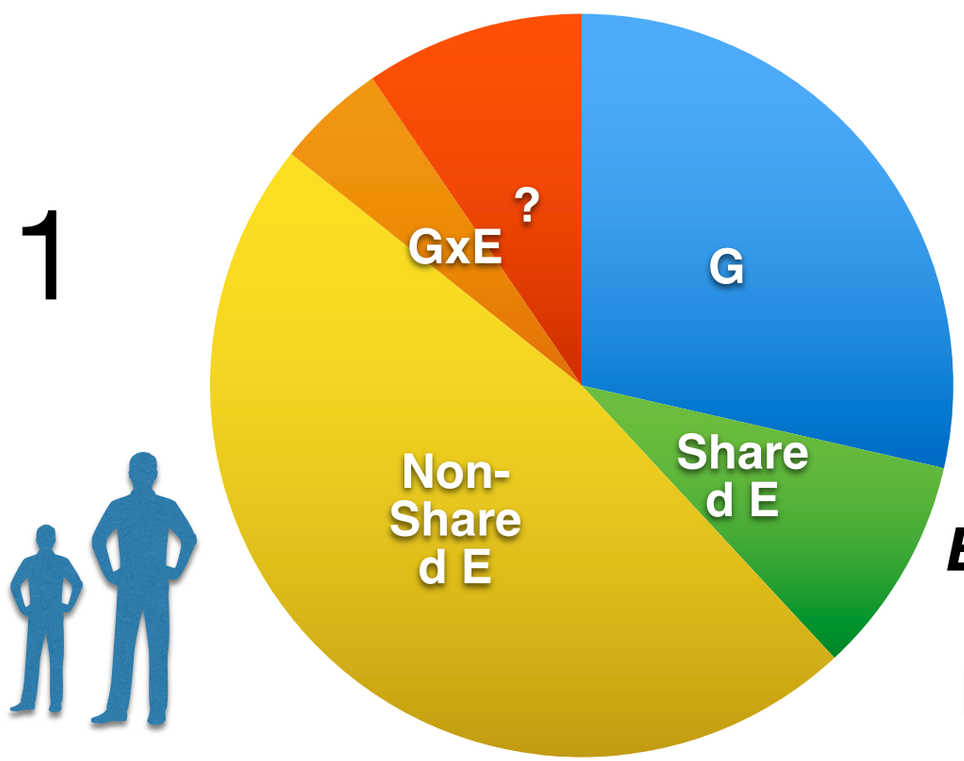
Type 2 Diabetes
Cancer
Alzheimer’s
Gene expression

Variants

Infectious agents
Diet + Nutrients
Pollutants
Drugs

Key data applications for precision medicine research:
(1) How much ***variation attributable to E*** in disease?
(2) What ***factors of the exposome*** are associated with disease?

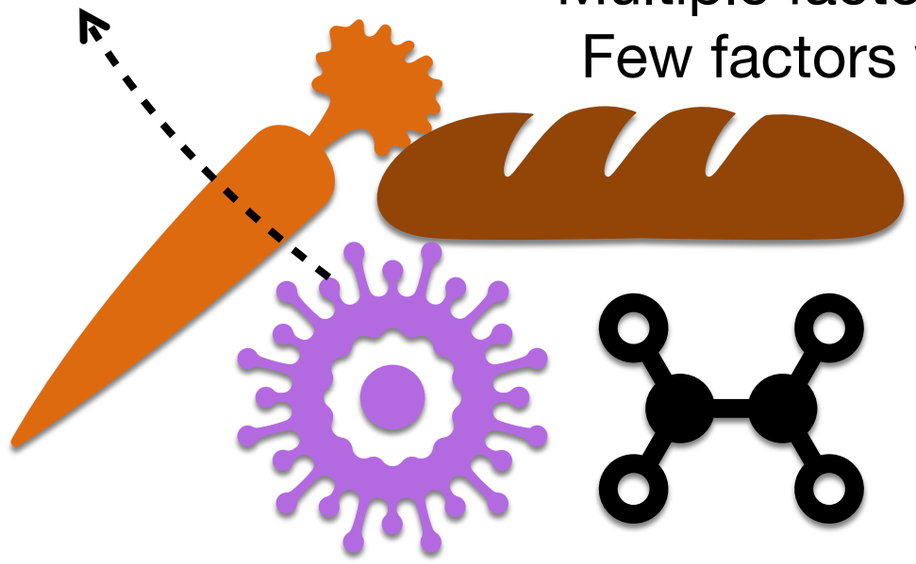
1



Larger the proportion (slice of pie):
More efficient discovery?

Exposure-wide association studies (ExWAS):
What factors are associated?
How do the exposures “add” up in aggregate?
Multiple factors of small effects?
Few factors with large effects?

2



?

Genomic and exposomic research are enhanced through large-scale cohort data, such as both ***GWAS*** and ***ExWAS!***

Diabetologia 2023
PLOS Biology 2021
PLOS Biology 2022
Nature Communications 2022
Nature Communications 2021
Diabetologia, 2021
Diabetes Care, 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**

Genomics and the *genome-wide association study:* an example of scalable, reproducible identification of genetic variation in disease

3,567 publications (as of 9/18/18)
71,673 *G-P* associations

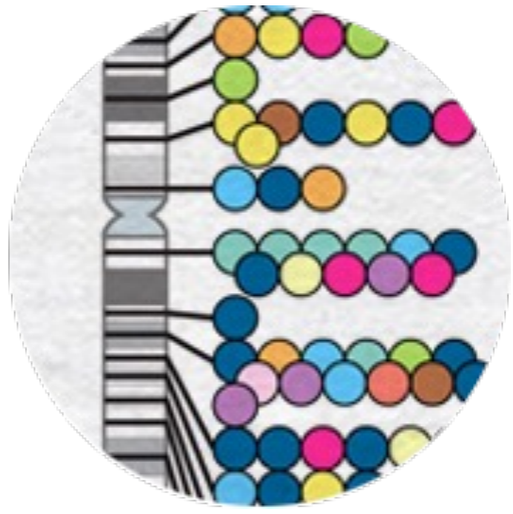
3,955 publications (as of 4/21/19)
136,287 *G-P* associations

4,493 publications (as of 3/10/20)
179,364 *G-P* associations

5,690 publications (as of 5/11/22)
372,752 *G-P* associations

6,422 publications (as of 7/5/23)
529,713 *G-P* associations

- Scaled for discovery
- Robust associations
- Negligible confounding
- Zero reverse causality
- *Little prediction capability*



https://www.ebi.ac.uk/gwas/

Abdellaoui et al, *AJHG* 2023

Possible to achieve translatable evidence with biobank scale data?

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

The Environment and Disease: Association or Causation?

by Sir Austin Bradford Hill CBE DSC FRCP(hon) FRS
(*Professor Emeritus of Medical Statistics, University of London*)

Proc R Soc Med 1965

"Bradford Hill" has been our "model" for assessment of *single* exposures in disease!

1.) Strength of association (high risk)

High odds ratio, risk ratios, variance explained…

2.) Consistency of association

Replicated in multiple cohorts and across groups

3.) Specificity of association

One exposure ~ one phenotype

4.)Temporality

Exposure comes before phenotype

5.) Biological gradient

Higher the exposure, the higher the risk

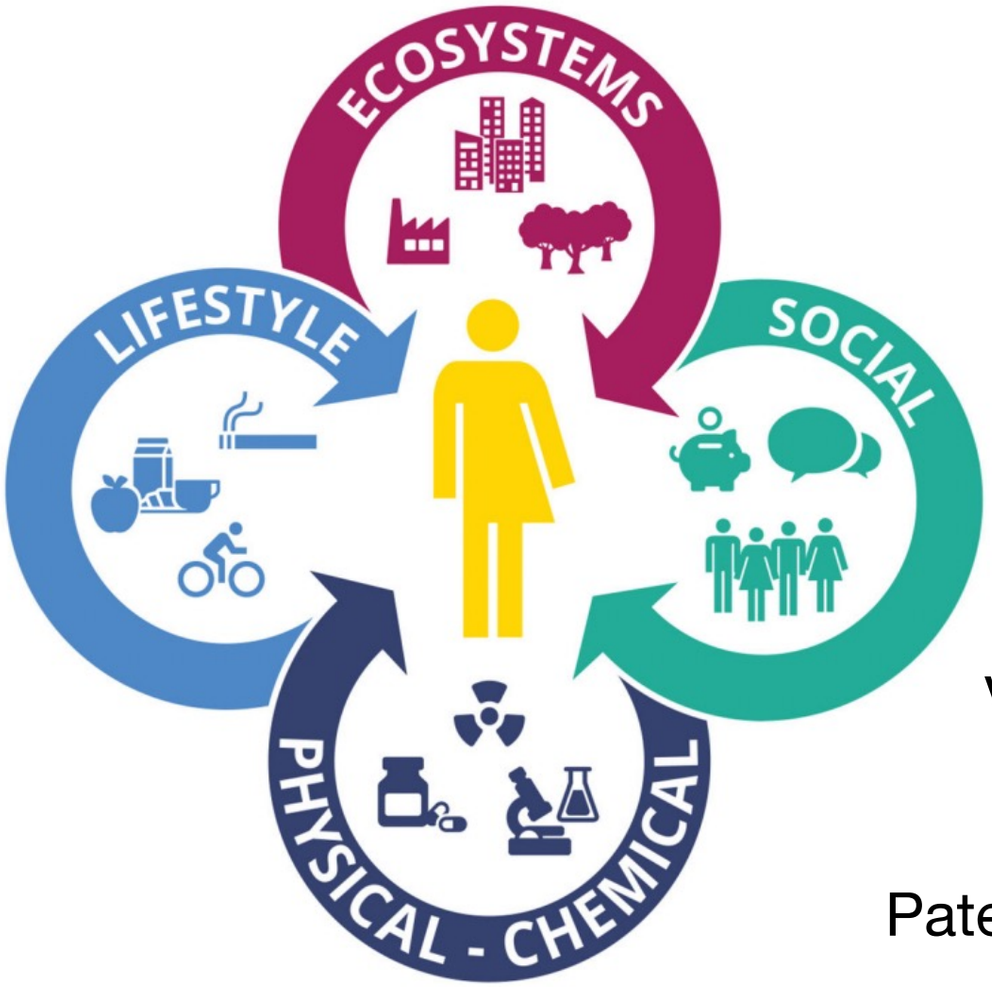
6.) Biological plausibility

Does the mechanism have some prior?

7.) Coherence

Can the association be reproduced experimentally

The exposome: systematic exposures across domains & modalities



Vermeulen R, Science 2020
Wild, Int J Epi 2012
Manrai et al., ARPH 2017
Patel and Ioannidis JAMA 2014
Ioannidis et al. STM 2009

ECOSYSTEMS
- Food outlets, alcohol outlets
- Built up environment and urban land uses
- Population density
- Walkability
- Green/Blue space

LIFESTYLE
- Physical activity
- Sleep behaviour
- Diet
- Drug use
- Smoking
- Alcohol use

SOCIAL
- Household income
- Inequality
- Social capital
- Social networks
- Cultural norms
- Cultural capital
- Psychological and mental stress

PHYSICAL - CHEMICAL
- Temperature/Humidity
- Electromagnetic Fields
- Ambient Light
- Odour & noise
- Point, line sources e.g. factories, ports
- Outdoor and indoor Air Pollution
- Agricultural activities, livestock
- Pollen/Mold/Fungus
- Pesticides
- Fragrance products (Musk, musk ketone)
- Flame Retardants (PBDEs)
- Persistent Organic Pollutants
- Plastics and plasticizers
- Food contaminants
- Soil contamination
- Drinking water contamination
- Groundwater contamination
- Surface water contamination
- Occupational exposures

Modalities of the ***exposome*** in the biobank records era are complex, time-dependent, and diverse in data type

Modality	**Type**	**Examples**
Targeted mass spec	Tabular; spectra	Lead; Cadmium; PFAS
Geospatial markers	Area-level; 2D spectra	Zipcode-level PM 2.5
Self-report questionnaire	Tabular; hierarchical	Nutritional recall
Untargeted mass spec	Tabular; spectra	Mass-charge ratio
Sensor-based behaviors	Tabular; spectra	Accelerometers

Patel et al, CEBP 2017
Manrai et al, ARPH 2017
Vermeulen et al, Science 2020

And the exposome is ***shared*** and ***non-shared***!

shared

Small particles in air pollution

non-shared

Behavior-related exposures

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

. . .

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

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

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

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

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

What is the “total” aggregate exposome, or the exposome “architecture” of phenotypes?

$$\sigma^2_E \; ?$$

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

$$\sigma^2_E = \sigma^2_{shared} + \sigma^2_{non\text{-}shared} \text{ (+ random chance)}$$

Lakhani et al., Nature Genetics 2019
See also: Rzhetsky et al Nature Comm 2019
Wang et al Nature Genetics 2017
Polubriaginof et al, Cell 2018

Creating cohorts with both G & E

Health insurance claims data to
*document the role of **genome** and **exposome** of patient phenotypes*

Weather

Air Pollution

Census SES

+

Health claims information

Disease (ICD9/ICD10),
procedures, drugs, labs
N ~ 45M

Lakhani et al., Nature Genetics 2019

Family relationships: a prerequisite to measure aggregate ***G*** *and* ***E*** *in* **501** ***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

We mapped ***13360*** ICD9 billing codes to ***1809*** ***PheWAS*** *(P)* codes (in addition to ***95*** Mendelian disorders)

CARDIOVASCULAR

hypertension (401)

cardiac dysrhythmias (427)

DIGESTIVE

irritable bowel syndrome (564.1)

ENDOCRINE

type 2 diabetes (250.1)

type 1 diabetes (250.2)

(and 11 more phenotype groups)

Denny, Bastarache, et al. 2013

Rzhetsky, White et al. 2013

We can estimate the ***proportion*** of fraternal and identical twins using ***opposite*** sex twin prevalence

$$h^2 = 2(r_{mz} - r_{dz})$$
$$c^2 = 2r_{dz} - r_{mz}$$

h^2 : narrow-sense **heritability**
c^2 : **shared environment**
r_{mz}: correlation of phenotype between ***identical*** twins
r_{dz}: correlation of phenotype between ***fraternal*** twins

Tetrachoric correlation to estimate r_{mz} & r_{dz}

… but we do **not** know the ***zygosity*** status of claimants…

But we do know:

Opposite sex twins: *all fraternal*
Same Sex twins 👯 : *mixture of identical and fraternal*

We can estimate the ***proportion*** of fraternal and identical twins using ***opposite*** sex twin prevalence

$$P(mz) \sim 1 - 2(N_{OS} / N_{all}) = \mathbf{0.26}$$

$$p(ss) = N_{ss} / N_{all} = \mathbf{0.63}$$

$$\boldsymbol{p} = P(mz|ss) = P(mz) / P(ss) = \mathbf{0.41}$$

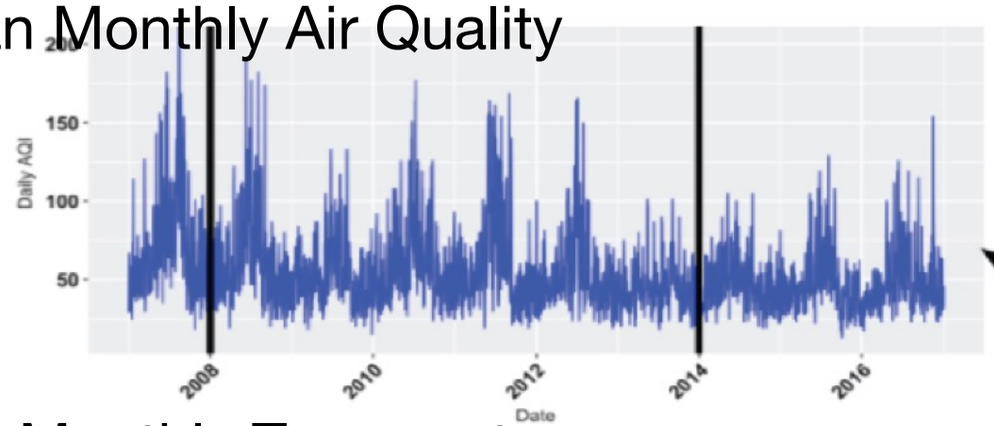
$$\mathbf{h^2} = 2/\boldsymbol{p}\ (r_{ss} - r_{os})$$

$$\mathbf{c^2} = (r_{os}(\boldsymbol{p}+1) - r_{ss}) \boldsymbol{/\ p}$$

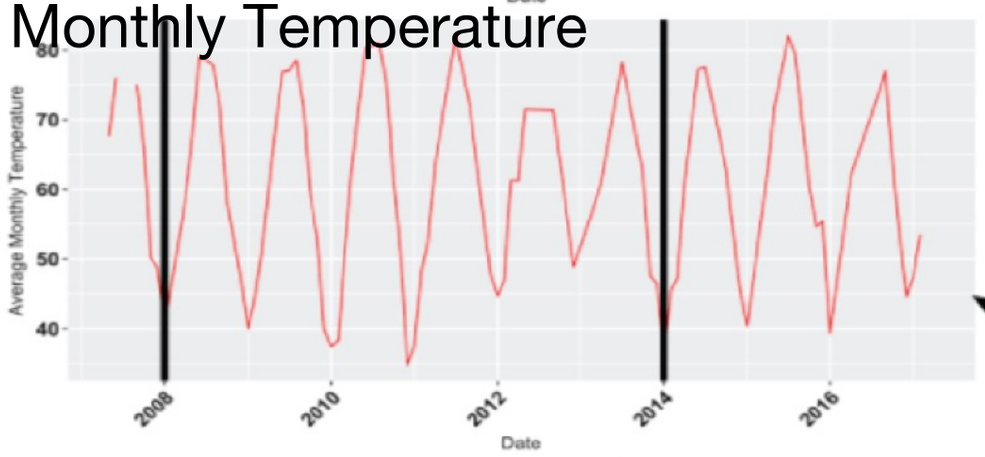
Weinberg, 1902
Benyamin, et al, 2005, 2006

Decoupling ***G*** from ***E*** in 560 ***P*** by integrating measured shared exposome (zipcode) and genome (twins)

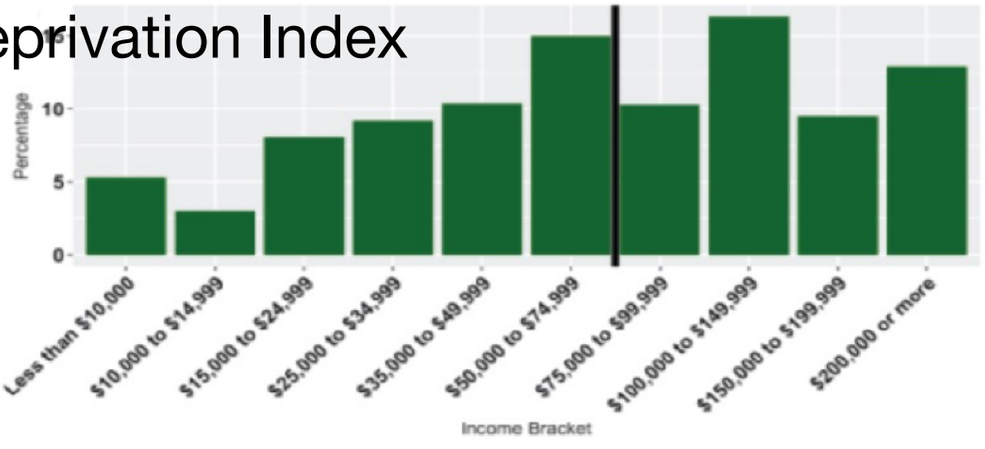
Median Monthly Air Quality

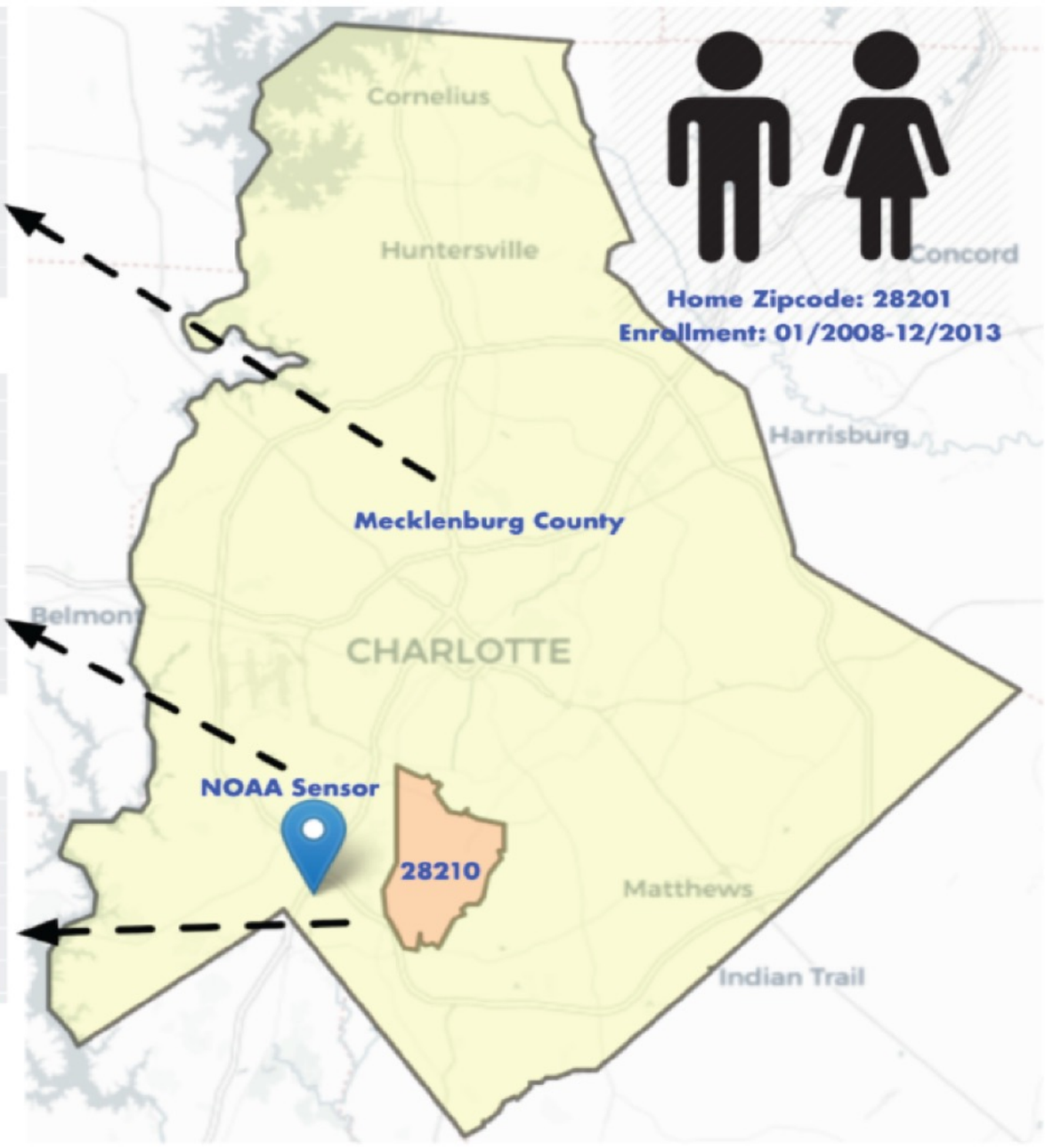


Median Monthly Temperature



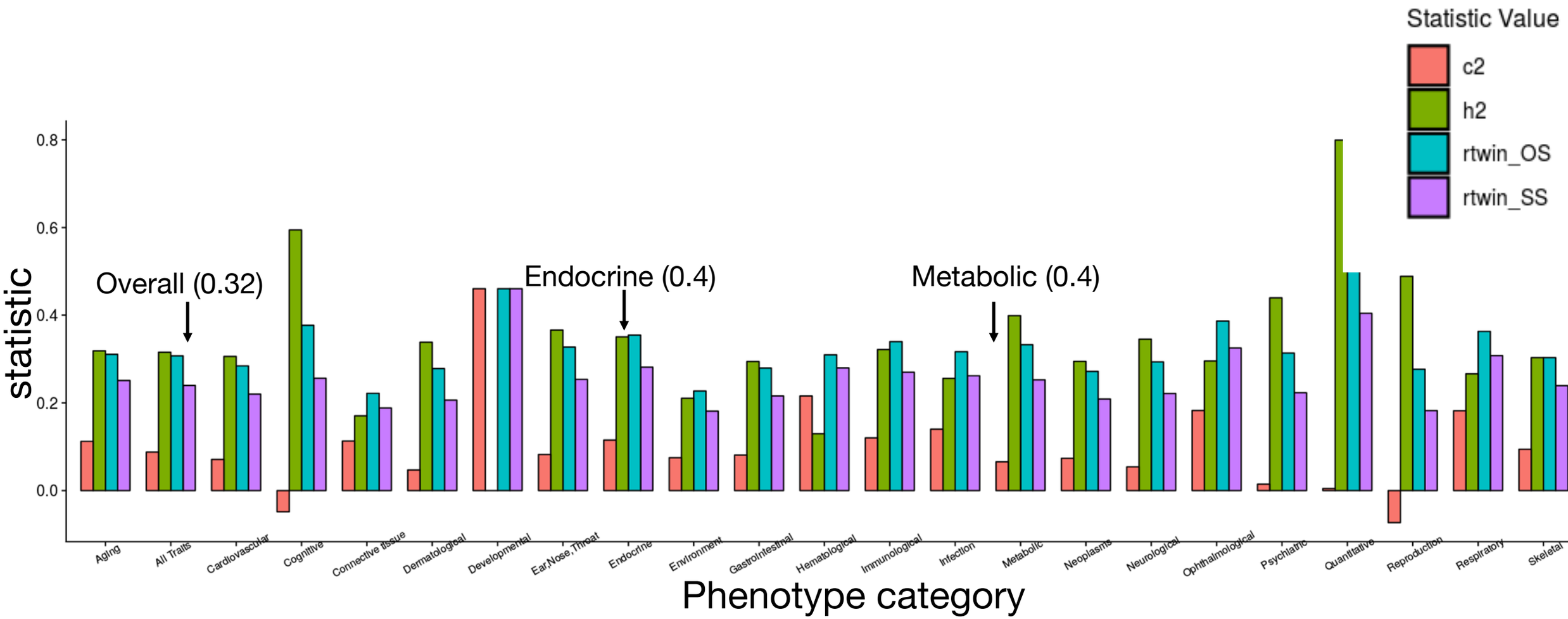
Sociodeprivation Index





Lakhani et al., Nature Genetics 2019

Patient cohorts in the "real-world" :
overall heritability (0.32) and shared environment (0.09):
modest (but reproducible) contributions of ***G and E***



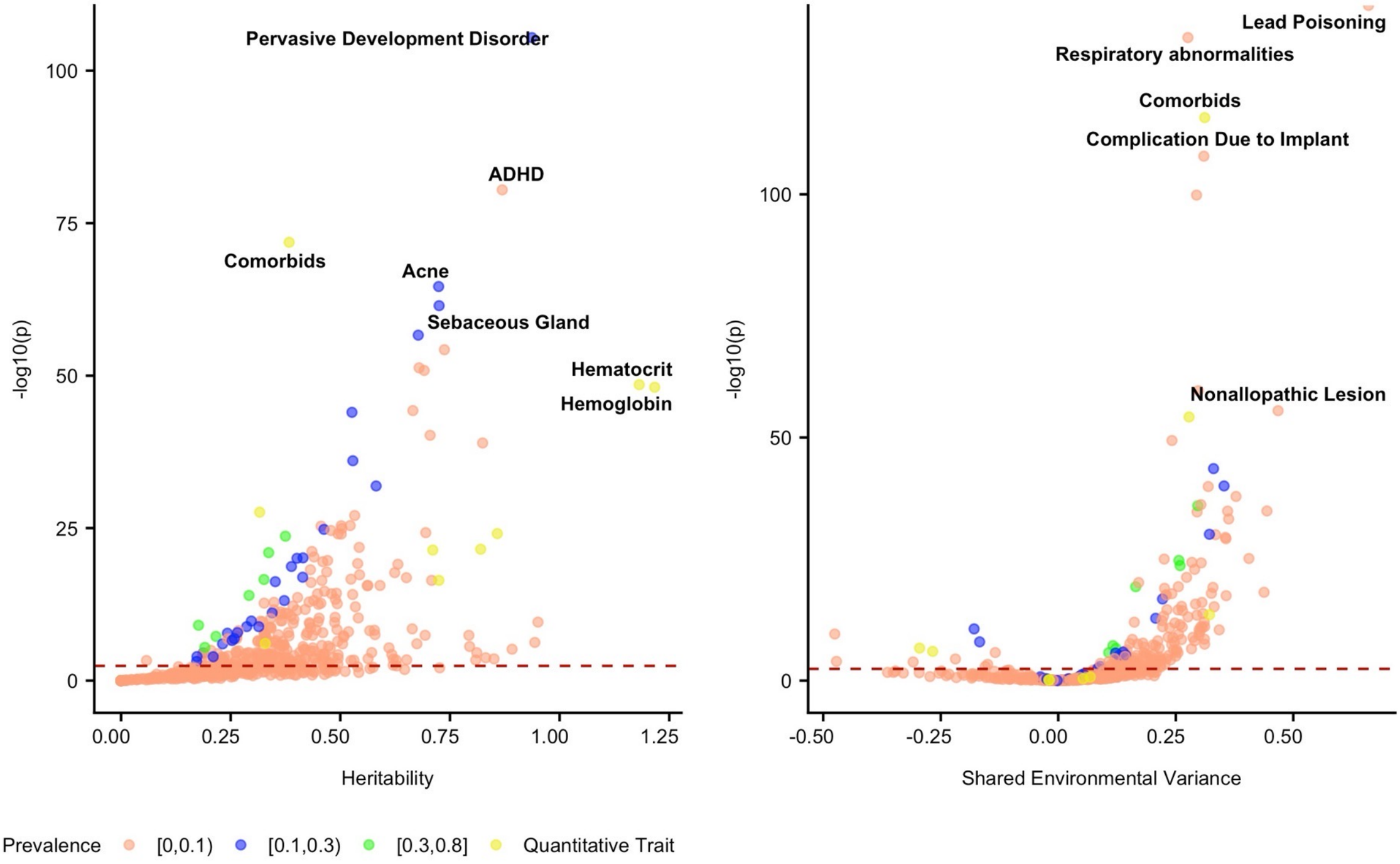
US-based, ages < 25

CaTCH: Claims analysis of Twin Correlation and Heritability

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

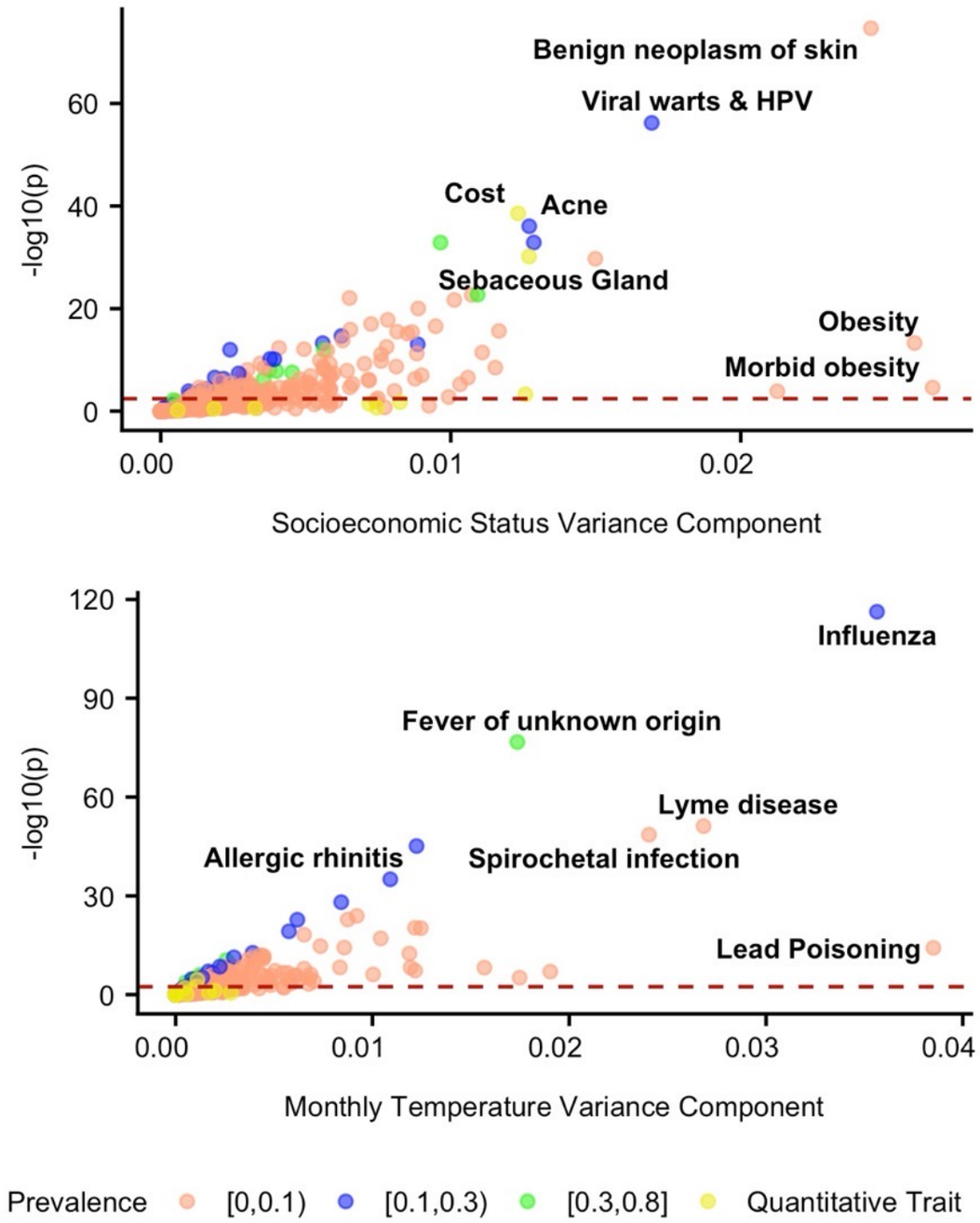
Lakhani et al., Nature Genetics 2019

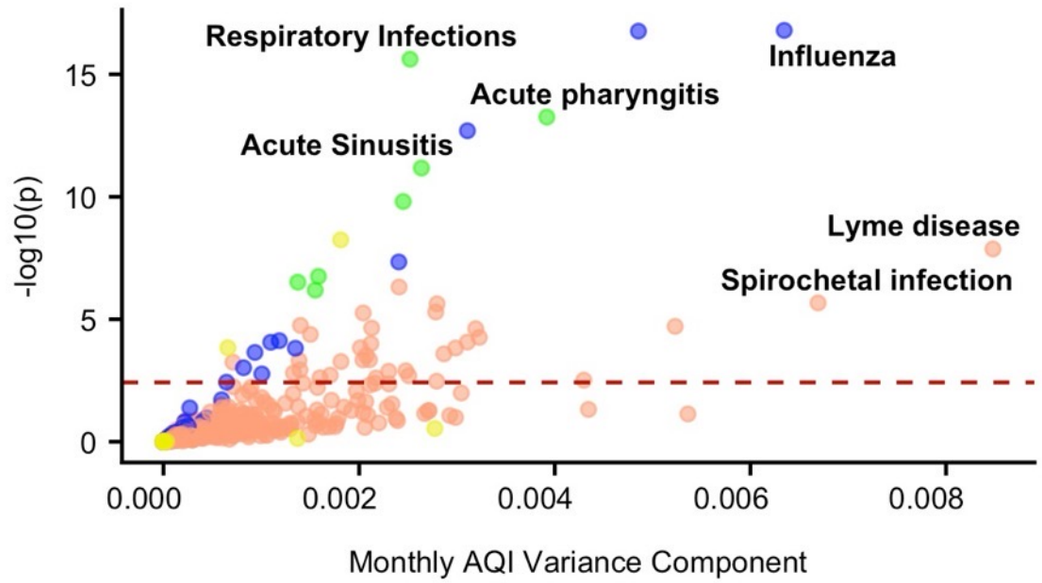
h^2 and c^2 estimates for ***560*** phenotypes versus statistical significance : 326/560 traits (>50%) have a heritable and 180/560 (32%) had a shared environment component!



Lakhani et al., Nature Genetics 2019

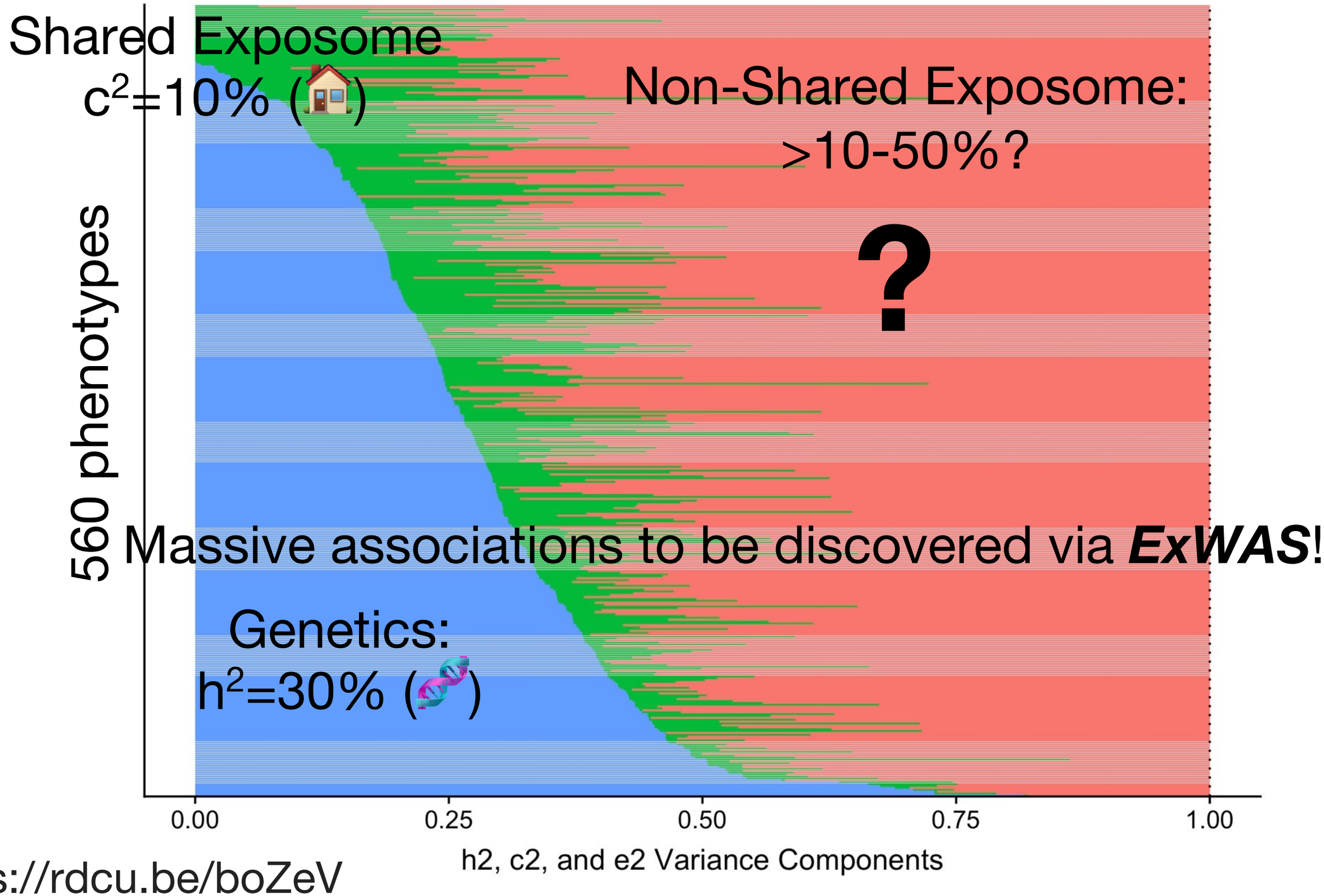
… air pollution, climate, and geocoded SES play a role in total shared environment, but cannot explain all of the variation!





Lakhani et al., Nature Genetics 2019

56K twins and 700K siblings in a massive health insurance cohort point to complex exposomic architecture in *P*



https://rdcu.be/boZeV

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

Lakhani et al., Nature Genetics 2019

Explaining the the missing 50-60% variation:
We are close to bringing ExWASs to practice,
but some challenges!

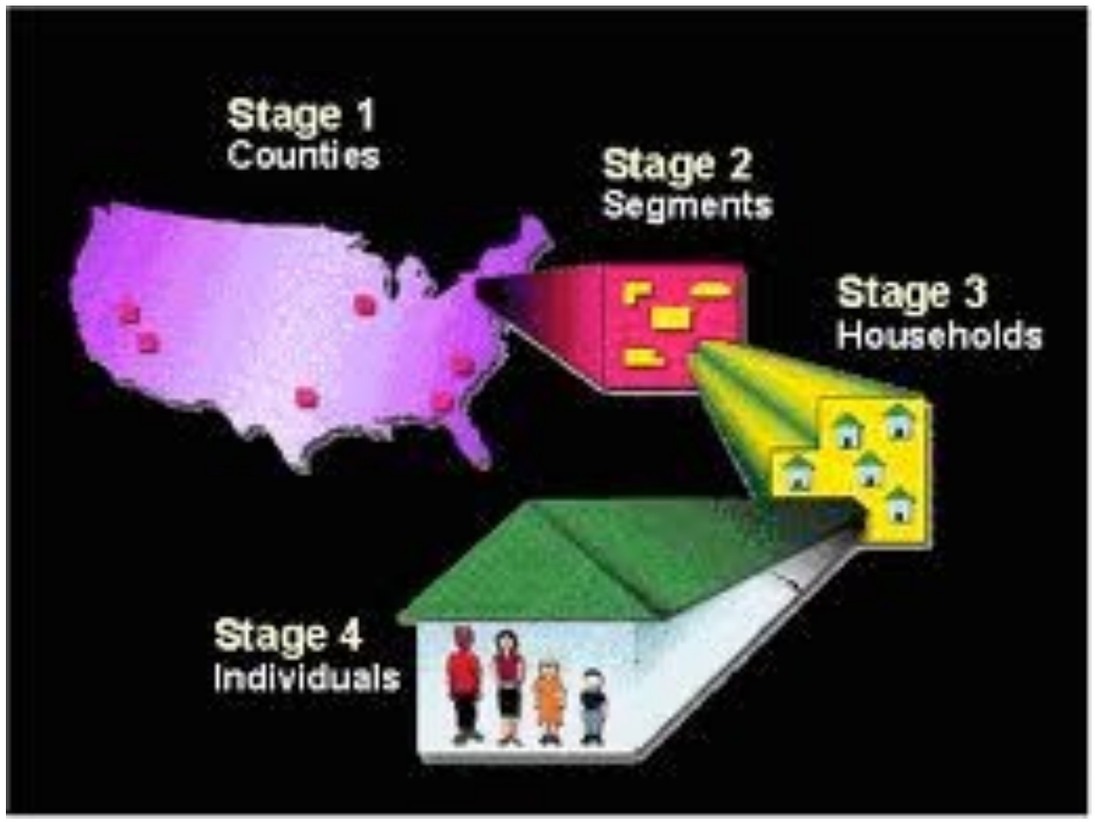
- ***What is the exposome****:* measurement technology and categories/domains of exposure
- ***Confounding & causality***: what factors to adjust for?
- ***Stability of exposures*** and longitudinal time to outcomes
- ***How large is the exposome:*** consideration of multiplicity (***false discoveries***)

Wild, 2005, 2012
Ioannidis , 2009
Rappaport and Smith, 2010, 2011
Buck-Louis and Sundaram 2012
Miller and Jones, 2014
Patel CJ and Ioannidis JPAI, 2014ab
Ioannidis, 2016
Manrai, 2017

Examples of ***exposome-driven*** discovery machinery,
or “***exposome-wide association studies***”

Gold standard for ***breadth*** of human exposure information: National Health and Nutrition Examination Survey[1]

since the 1960s
now biannual: 1999 onwards
10,000 participants per survey



>250 exposures (serum + urine)
GWAS chip

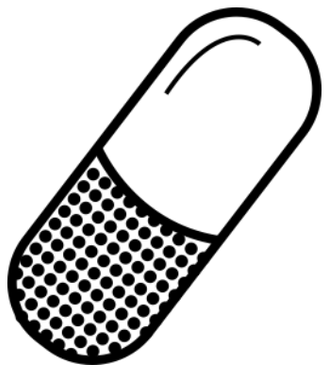
>85 quantitative clinical traits (e.g., serum glucose, lipids, body mass index)

Death index linkage (cause of death)

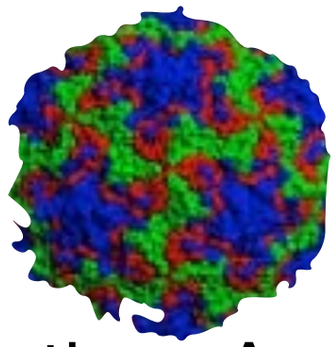
[1] http://www.cdc.gov/nchs/nhanes.htm

Gold standard for ***breadth*** of exposure & behavior data:
National Health and Nutrition Examination Survey

Nutrients and Vitamins
vitamin D, carotenes



Drugs
statins; aspirin



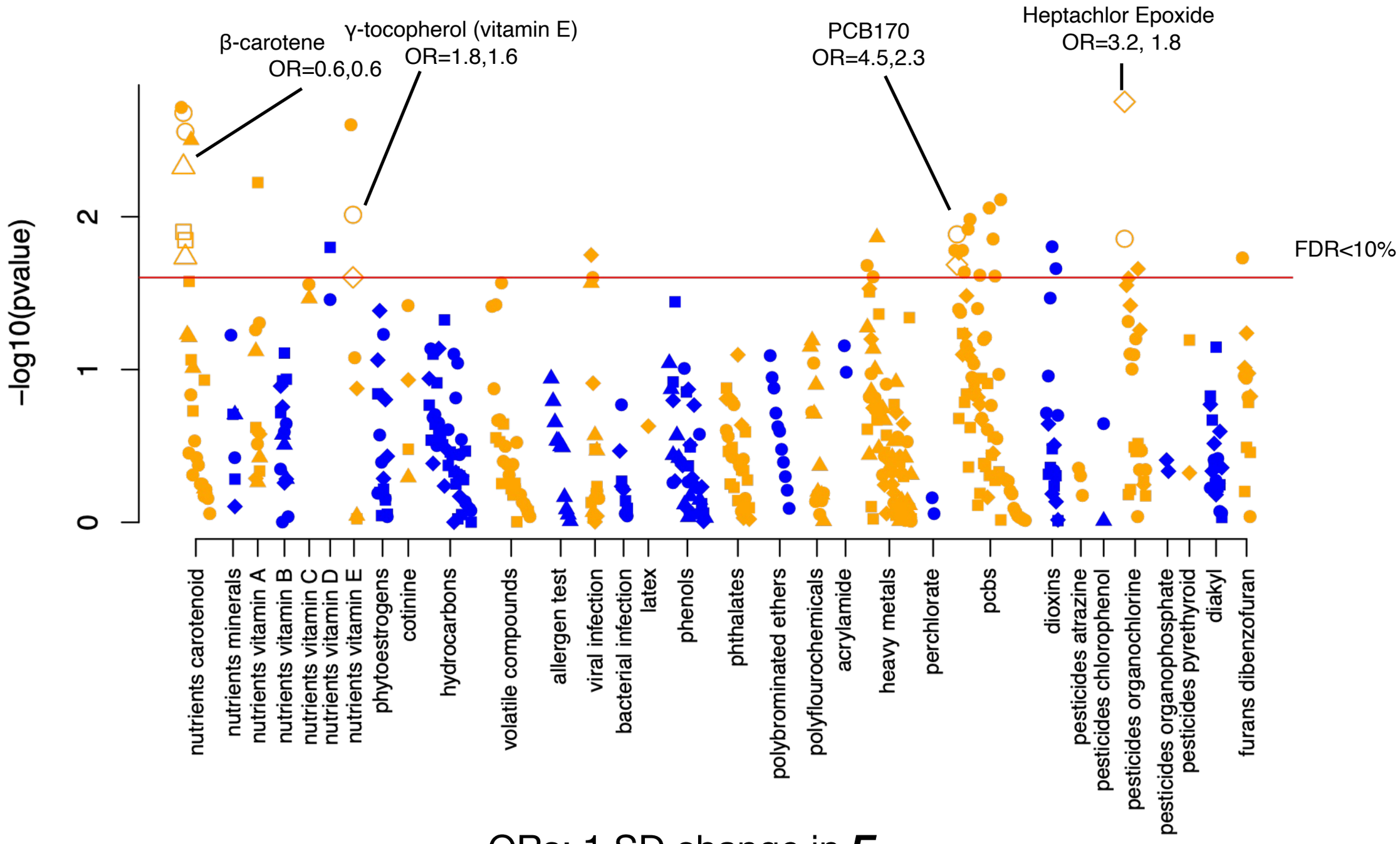
Infectious Agents
hepatitis, HIV, Staph. aureus

Plastics and consumables
phthalates, bisphenol A

Pesticides and pollutants
atrazine; cadmium; hydrocarbons

Physical Activity
e.g., steps

Going “exposome-wide” **in Type 2 Diabetes**: Serum nutrients and persistent chemicals associated with FBG > 125 mg/dL

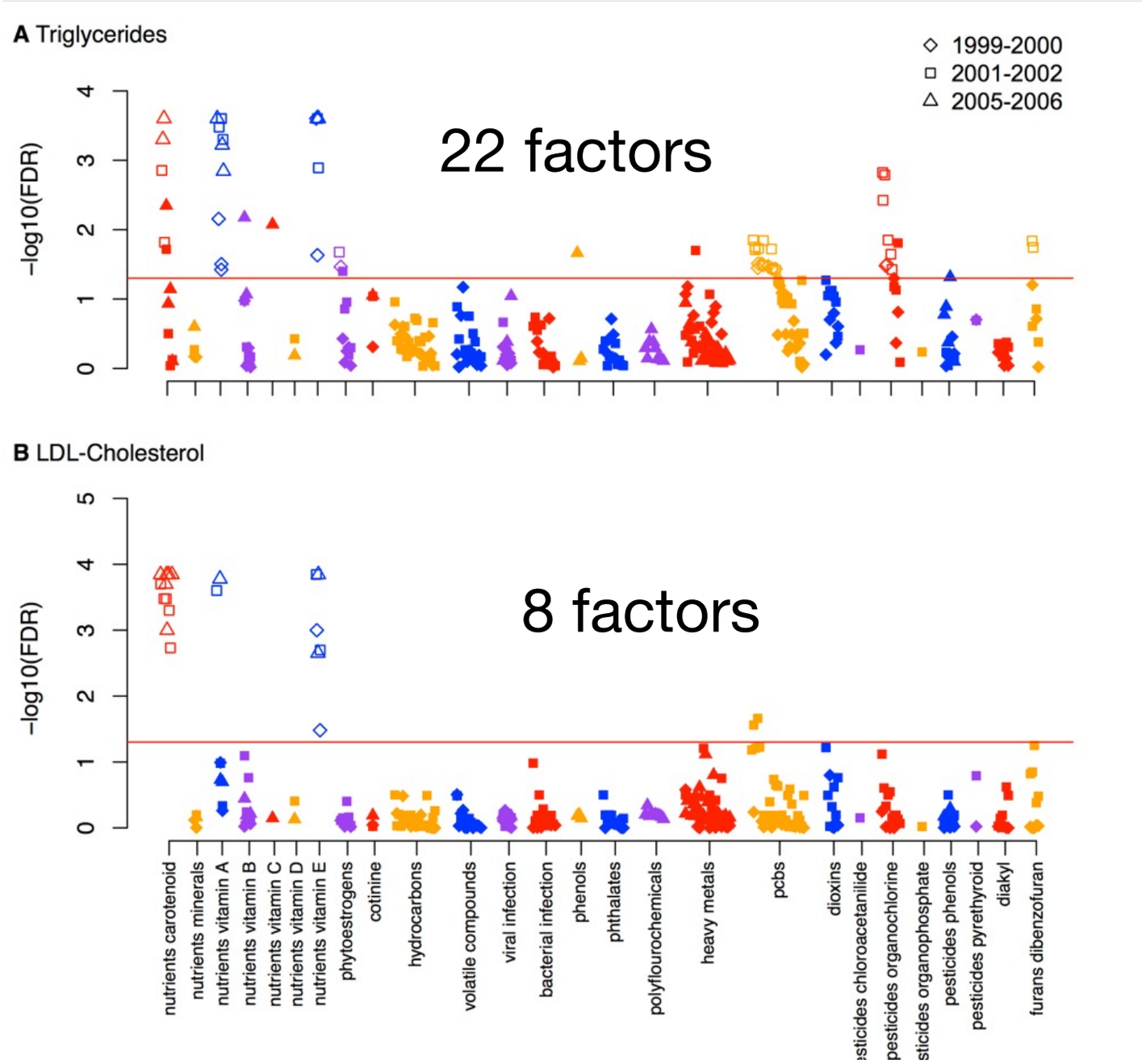


ORs: 1 SD change in ***E***

N=100-3000 per survey (4 surveys)

PLOS ONE 2010

EWAS in triglycerides identifies 22 ***E*** associations (11%); however, fewer ***E*** (4%) in LDL-C



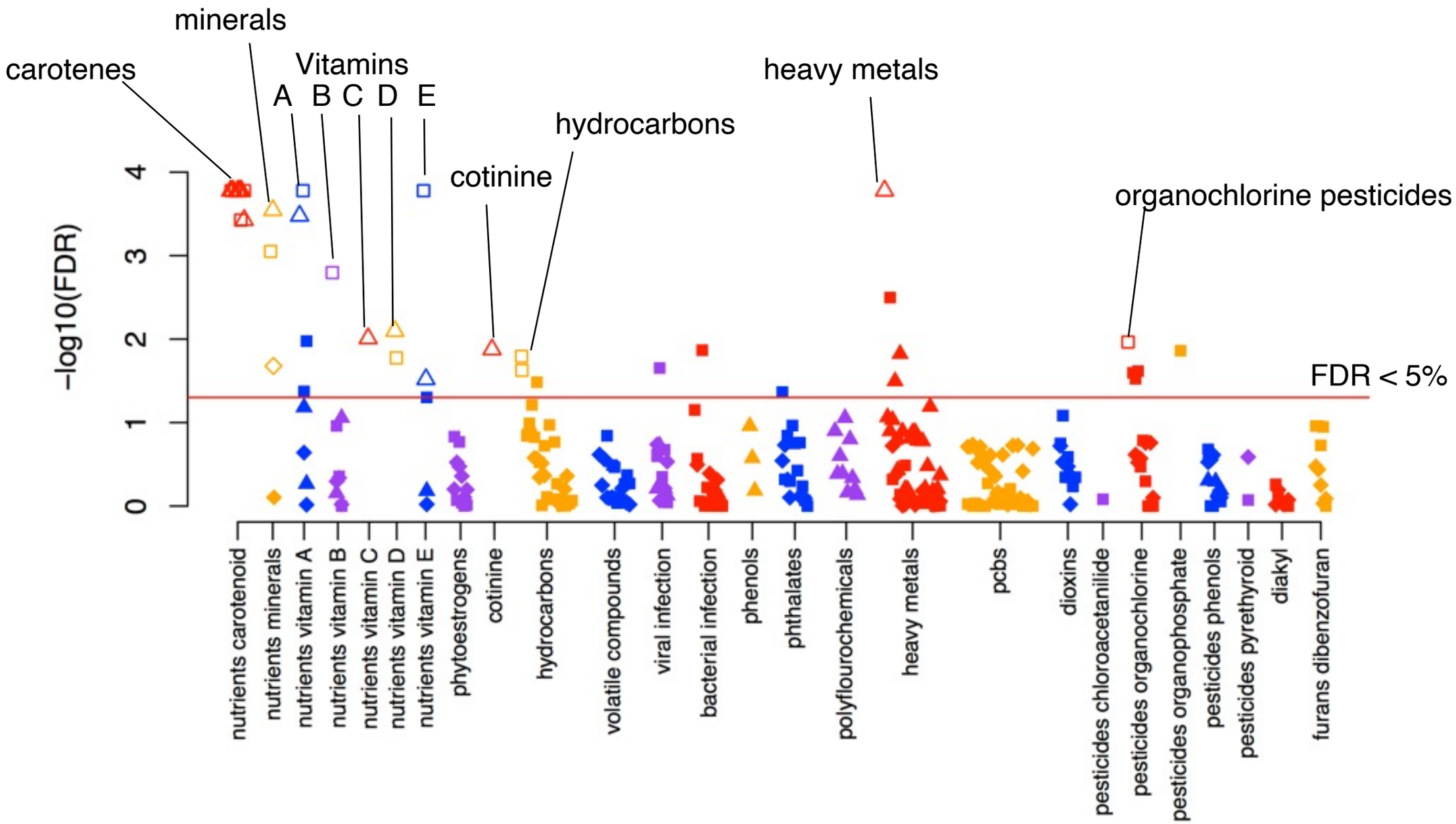
organochlorine pesticides
polychlorinated biphenyls
carotenoids
vitamin E
vitamin A

carotenoids
vitamin E
vitamin A

1-15 mg/dL
R^2 ~ 15, 2%

Int J Epidem 2012.

Broad spectrum of serum nutrients, persistent pollutants, and behavior (cotinine) associated with HDL-C **(17 out of 188 [9%])**

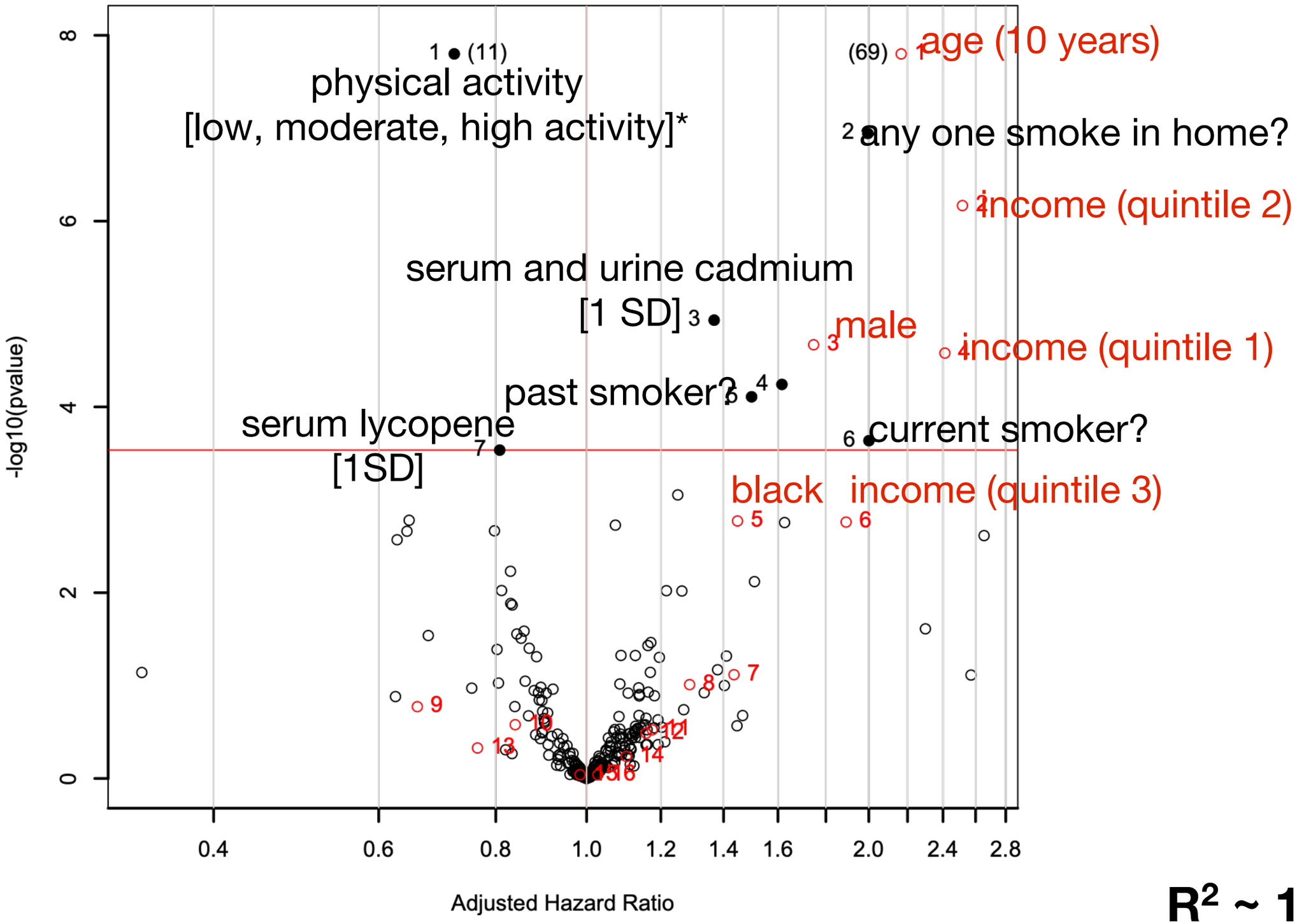


$\log_{10}$(HDL-C)
adjusted for BMI, SES, ethnicity, age, age^2, sex
N=1000-3000

1-5 mg/dL
R^2 ~ 15%

Int J Epidem. 2012

ExWAS identifies factors associated with ***all-cause mortality***:
HR vs. -log10(pvalue) of 253 associations



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

R^2 ~ 14%
(2%)

Researching the exposome: where has exposome-wide taken us in 12-15 years?

National Library of Medicine
National Center for Biotechnology Information

Log in

PubMed.gov

exposome or exposome-wide or environment-wide

Search

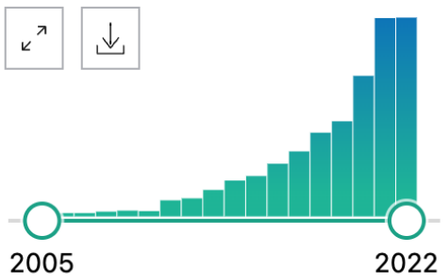
Advanced Create alert Create RSS User Guide

Save Email Send to Sorted by: Most recent Display options

MY NCBI FILTERS

1,639 results Page 1 of 164

RESULTS BY YEAR



TEXT AVAILABILITY

- Abstract
- Free full text
- Full text

ARTICLE ATTRIBUTE

- Associated data

Did you mean ***expo one or exposome wide or environment wide*** (81,054 results)?

1 Cite Share

Understanding the Chemical **Exposome** During Fetal Development and Early Childhood: A Review.

Krausová M, Braun D, Buerki-Thurnherr T, Gundacker C, Schernhammer E, Wisgrill L, Warth B.

Annu Rev Pharmacol Toxicol. 2022 Oct 6. doi: 10.1146/annurev-pharmtox-051922-113350. Online ahead of print.

PMID: 36202091 Review.

Infants are exposed to a multitude of environmental factors, collectively referred to as the **exposome**. The chemical **exposome** can be summarized as the sum of all xenobiotics that humans are exposed to throughout a lifetime. ...Several recommendations to advance our u ...

Paperpile

2 Cite Share

Benefits of topical hyaluronic acid for skin quality and signs of skin aging: from literature review to clinical evidence.

Bravo B, Correia P, Junior JEG, Sant'Anna B, Kerob D.

Dermatol Ther. 2022 Oct 6:e15903. doi: 10.1111/dth.15903. Online ahead of print.

PMID: 36200921 Review.

Diverse association sizes/variance for ~300 *E* factors illuminates the broad implications for risk and biology

A Nutrient-Wide Association Study on Blood Pressure

Ioanna Tzoulaki, PhD;* Chirag J. Patel, PhD;* Tomonori Okamura, MD, PhD; Queenie Chan, PhD; Ian J. Brown, PhD; Katsuyuki Miura, MD, PhD; Hirotsugu Ueshima, MD, PhD; Liancheng Zhao, MD; Linda Van Horn, PhD; Martha L. Daviglus, MD, PhD; Jeremiah Stamler, MD; Atul J. Butte, MD, PhD; John P.A. Ioannidis, MD, DSc; Paul Elliott, MB BS, PhD

Circulation 2012
Betas: 0.9-1.3 per 1 SD
82 *E*
R^2 (SBP): <1%

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* and *P*
IJE 2017
R^2: 1%

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
R^2 (triglycerides): 10%
R^2 (LDL): 2%
R^2 (HDL): 15%
249 *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]*

IJE 2013
188 *E*
HRs: 0.7-2.8 (per 1SD)
Nagelkerke R^2: 2%

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]✉

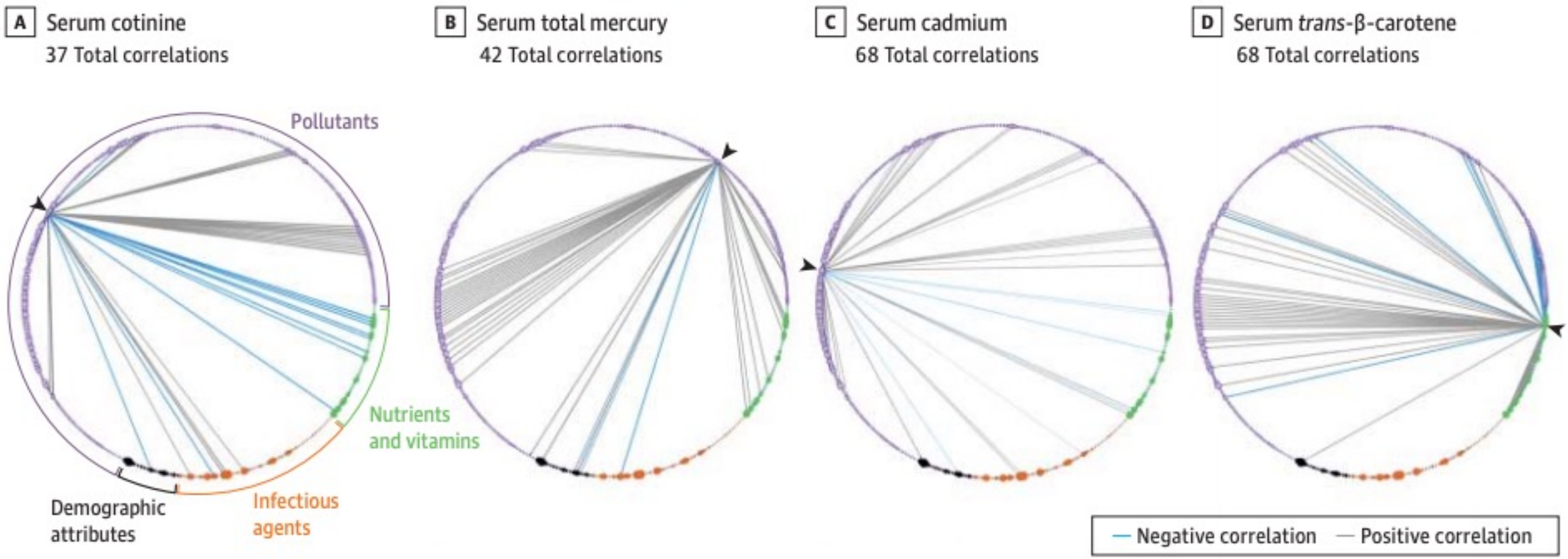
374 *E*
Nature Communications 2020
HRs: 0.9-1.3 per 1 SD

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,*]

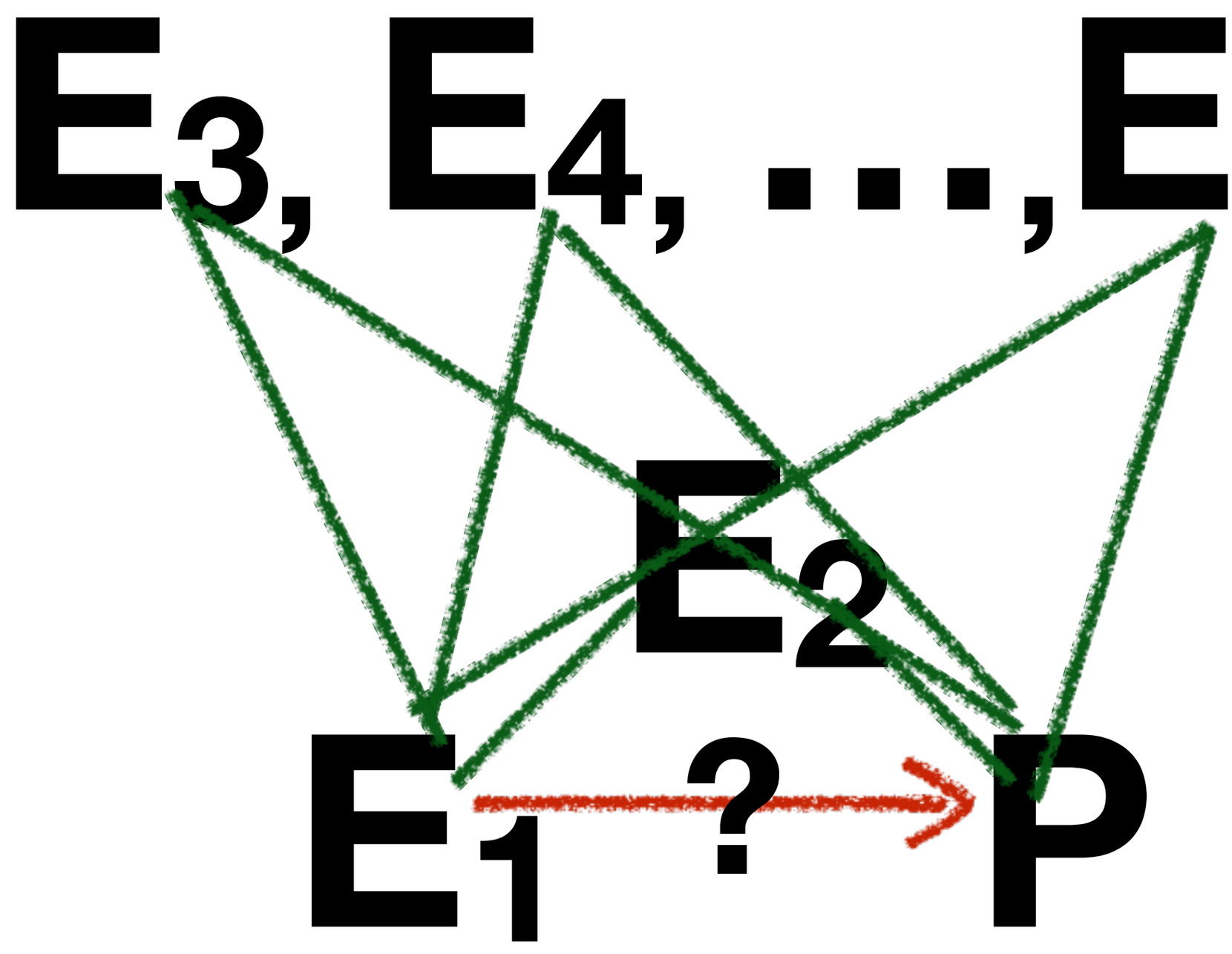
128 *E*
Env Int 2018

Correlation structure between *E* factors:
Correlation “globes” for 4 factors is dense but modest in overall association (average correlation of 0.3)



Pac Symp Biocomput. 2015
JECH. 2015
JAMA 2014

In massive ***non-genetic*** data:
The potential for confounding can be immense!



MEDICINE

Big data meets public health

Human well-being could benefit from large-scale data if large-scale noise is minimized

By **Muin J. Khoury**[1,2] *and*
John P. A. Ioannidis[3]

In 1854, as cholera swept through London, John Snow, the father of modern epidemiology, painstakingly recorded the locations of affected homes. After long, laborious work, he implicated the Broad Street water pump as the source of the outbreak, even without knowing that a *Vibrio* organism caused cholera. "Today, Snow might have crunched Global Positioning System information and disease prevalence data, solving the problem within hours" (*1*). That is the potential impact of "Big Data" on the public's health. But the promise of Big Data is also accompanied by claims that "the scientific method itself is becoming obsolete" (*2*), as next-generation computers, such as IBM's Watson (*3*), sift through the digital world to provide predictive models based on massive information. Separating the true signal from the gigantic amount of noise is neither easy nor straightforward, but it is a challenge that must be tackled if information is ever to be translated into societal well-being.

The term "Big Data" refers to volumes of large, complex, linkable information (*4*). Beyond genomics and other "omic" fields, Big Data includes medical, environmental, financial, geographic, and social media information. Most of this digital information was unavailable a decade ago. This swell of data will continue to grow, stoked by sources that are currently unimaginable. Big Data stands to improve health by providing insights into the causes and outcomes of disease, better

From validity to utility. Big Data can improve tracking and response to infectious disease outbreaks, discovery of early warning signals of disease, and development of diagnostic tests and therapeutics.

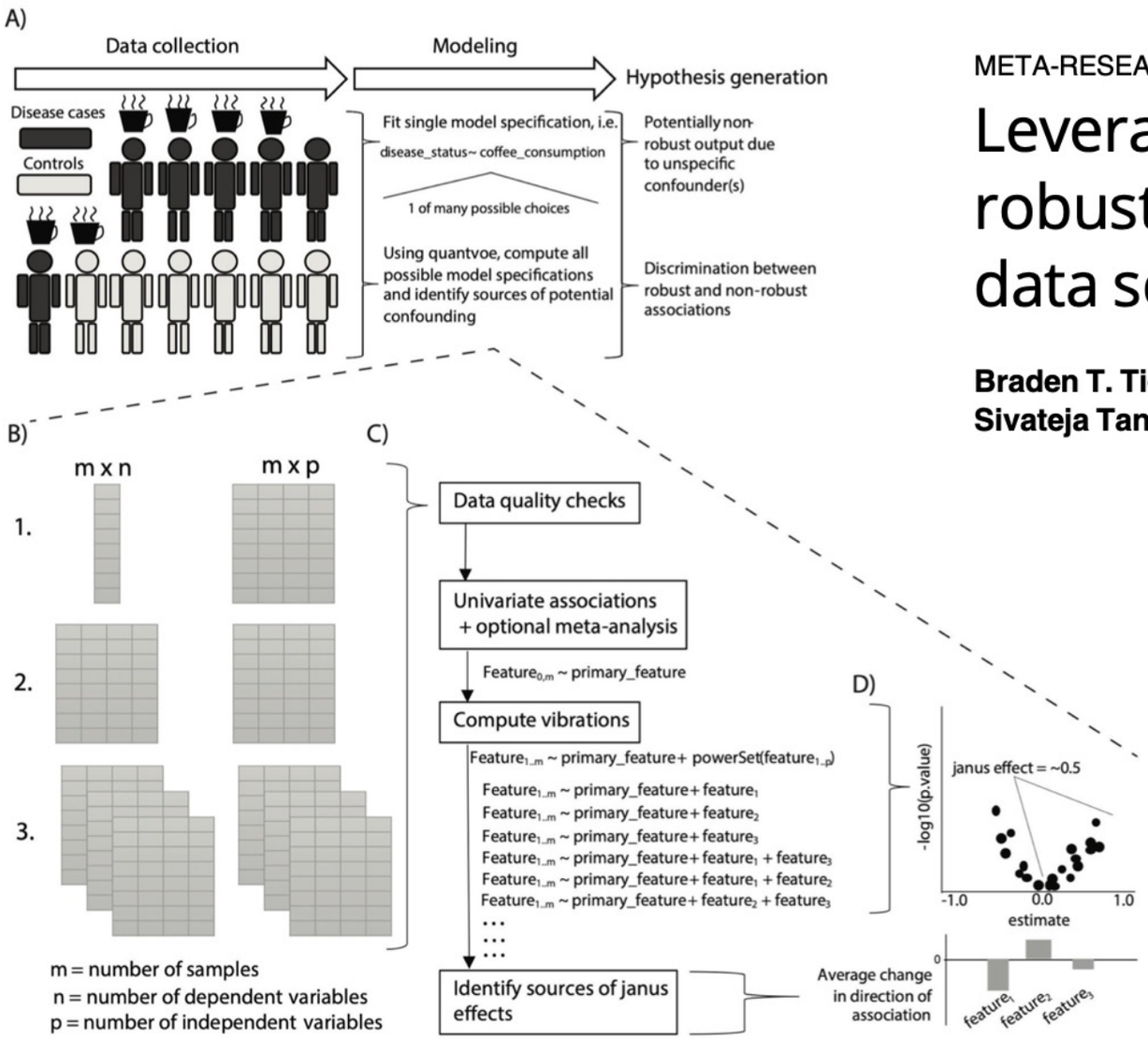
For nongenomic associations, false alarms due to confounding variables or other biases are possible even with very large-scale studies, extensive replication, and very strong signals (*9*). Big Data's strength is in finding associations, not in showing whether these associations have meaning. Finding a signal is only the first step.

Even John Snow needed to start with a plausible hypothesis to know where to look, i.e., choose what data to examine. If all he had was massive amounts of data, he might well have ended up with a correlation as spurious as the honey bee–marijuana connection. Crucially, Snow "did the experiment." He removed the handle from the water pump and dramatically reduced the spread of cholera, thus moving from correlation to causation and effective intervention.

How can we improve the potential for Big Data to improve health and prevent disease? One priority is that a stronger epidemiological foundation is needed. Big Data analysis is currently largely based on convenient samples of people or information available on the Internet. When associations are probed between perfectly measured data (e.g., a genome sequence) and poorly measured data (e.g., administrative claims health data), research accuracy is dictated by the weakest link. Big Data are observational in nature and are fraught with many biases such as selection, confounding variables, and lack of generalizability. Big Data analysis may be embedded in epidemiologically well-characterized and representative populations. This epidemiologic approach has served the genomics

Science 2015

QuantVoE: scaling up sensitivity analyses to test robustness of modeling scenarios (is it enough to adjust for *a priori* variables?)



META-RESEARCH ARTICLE

Leveraging vibration of effects analysis for robust discovery in observational biomedical data science

Braden T. Tierney[1,2,3,4], Elizabeth Anderson[1], Yingxuan Tan[1], Kajal Claypool[1], Sivateja Tangirala[1,5], Aleksandar D. Kostic[2,3,4], Arjun K. Manrai[1,6], Chirag J. Patel[1]*

Tierney et al, *PLOS Biology* 2021

https://github.com/chiragjp/quantvoe

See also: Tierney et al, *PLOS Biology* 2022
Tierney et al., *Nature Communications* 2021

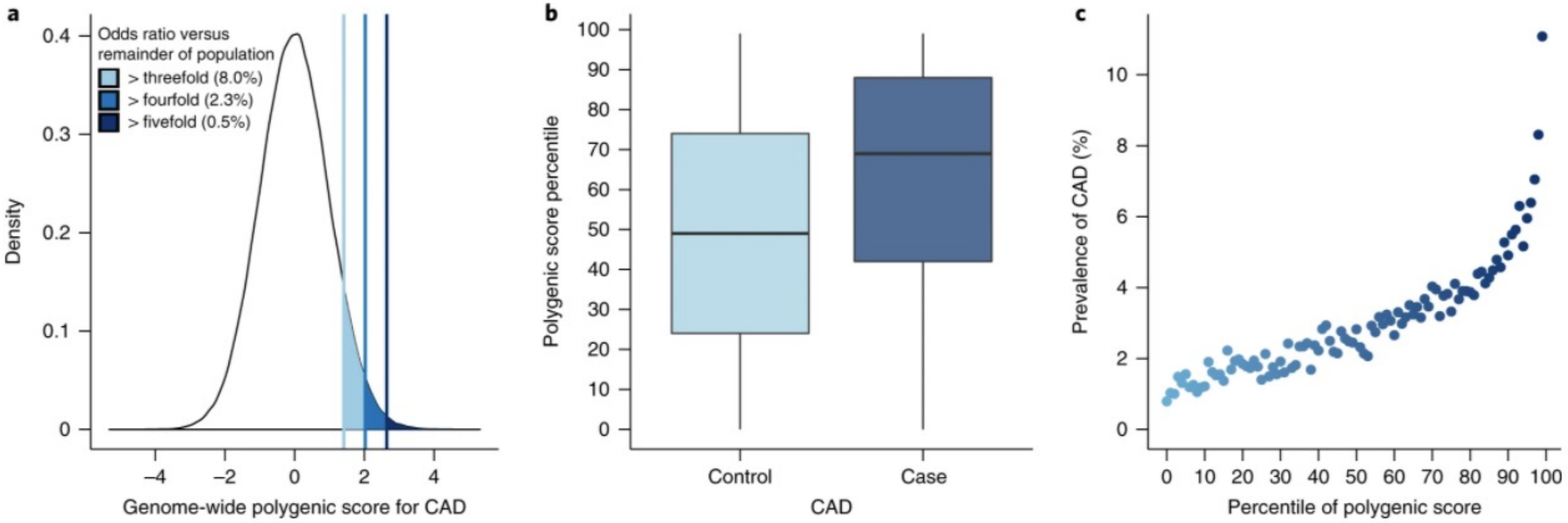
What evidence is needed to translate genomics and exposomics to the bedside?

The ***polygenic risk score*** (PRS), or ***genome wide predictive*** score (GPS) has emerged as a way measure cumulative genetic "burden"

- Are GWAS variants clinically relevant?
- Any one variant may not be (odds ratios are small)
- In contrast, **polygenic risk scores:**
  - Summarize ***additive genetic risk*** for disease in a ***time-invariant*** way
  - Are the sum of the association sizes ***(e.g., the odds ratios)*** for each variant for an individual

Khera et al. *Nature Genetics 2018*

tification of coronary artery disease according to



a, Distribution of GPS_{CAD} in the UK Biobank testing dataset (n = 288,978). The x axis represents GPS_{CAD}, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation. Shading reflects the proportion of the population with three-, four-, and fivefold increased risk versus the remainder of the population. The odds ratio was assessed in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. **b**, GPS_{CAD} percentile among CAD cases versus controls in the UK Biobank testing dataset. Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range, and the whiskers reflect the maximum and minimum values within each grouping. **c**, Prevalence of CAD according to 100 groups of the testing dataset binned according to the percentile of the GPS_{CAD}.

Odds of CAD > 5 fold in top 0.5% of population

Khera et al. *Nature Genetics 2018*

Building a ***P***oly-e**X**posure Risk **S**core (***PXS***): UK Biobank, 111 modifiable/non-modifiable exposures

N=111

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

Comparisons of Polyexposure, Polygenic, and Clinical Risk Scores in Risk Prediction of Type 2 Diabetes

Diabetes Care 2021;44:935–943 | https://doi.org/10.2337/dc20-2049

Yixuan He,[1,2] Chirag M. Lakhani,[2] Danielle Rasooly,[2,3] Arjun K. Manrai,[2,3] Ioanna Tzoulaki,[4,5] and Chirag J. Patel[2]

Diabetes Care 2021
UK Biobank

Questionnaire-Based Polyexposure Assessment Outperforms Polygenic Scores for Classification of Type 2 Diabetes in a Multiancestry Cohort

https://doi.org/10.2337/dc22-0295

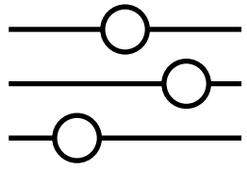
Farida S. Akhtari,[1,2] Dillon Lloyd,[1] Adam Burkholder,[3] Xiaoran Tong,[1] John S. House,[1] Eunice Y. Lee,[1] John Buse,[4] Shepherd H. Schurman,[2] David C. Fargo,[3] Charles P. Schmitt,[5] Janet Hall,[2] and Alison A. Motsinger-Reif[1]

Diabetes Care 2022

Personalized Environment and Genes (PEGS) cohort

Building a ***P***oly-e**X**posure Risk **S**core (***PXS***): UK Biobank, 111 modifiable/non-modifiable exposures

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

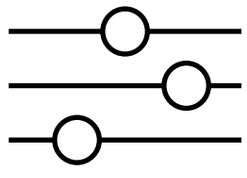


Filter & Select
XWAS
Lasso
P value thresholds

Diabetes Care 2021

Building a ***P***oly-e**X**posure Risk **S**core (***PXS***): UK Biobank, 111 modifiable/non-modifiable exposures

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

PRS and ***PXS (Poly eXposure Score):***
C-index increases that may be complementary
(but both much less than simple demographics and clinical factors)

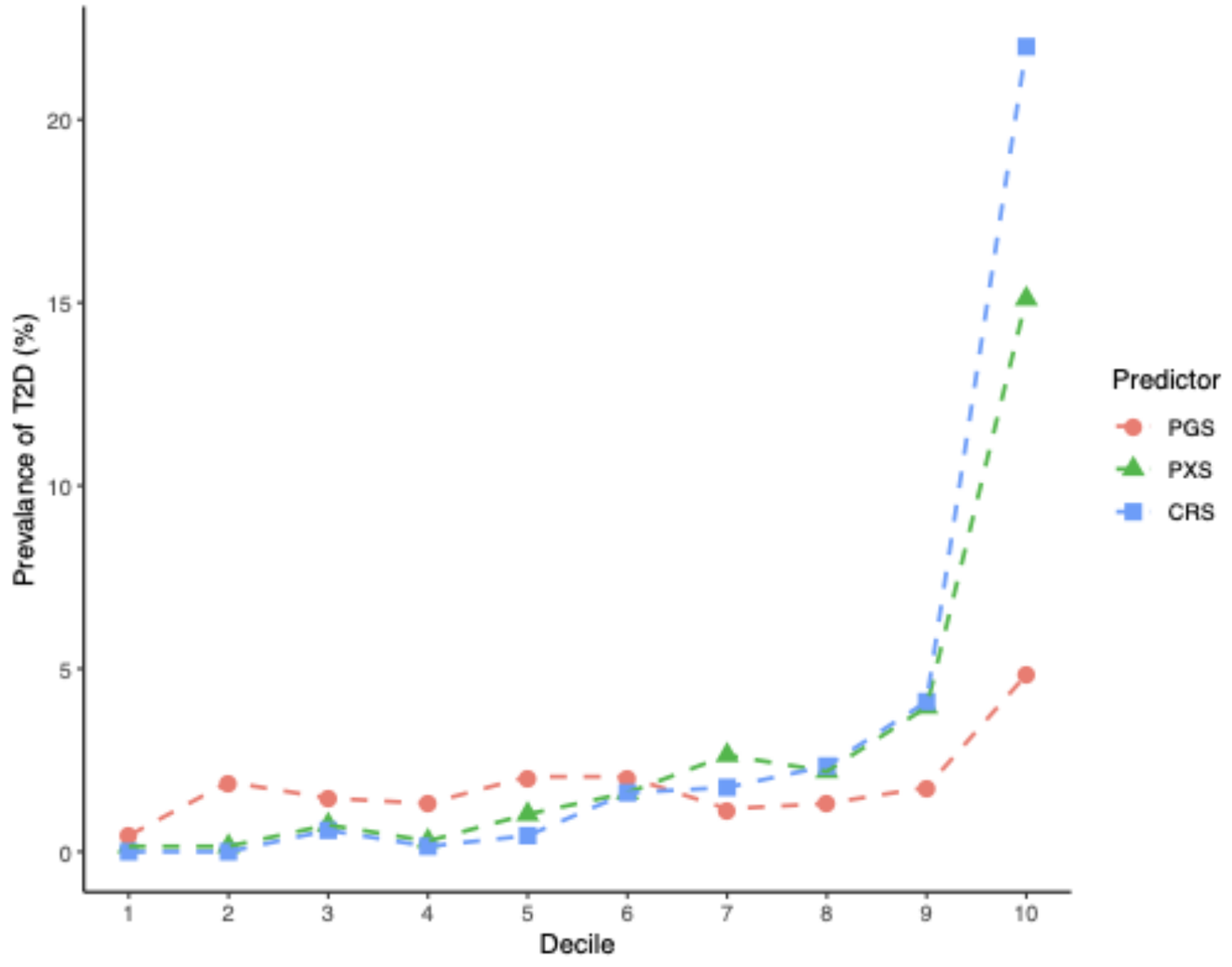
	C-Statistic (95% CI)		
	All	**Male**	**Female**
N	68299	32657	35642
# of Events	1281	844	437
Sex+Age	0.670 (0.656, 0.684)	0.629 (0.612, 0.646)	0.637 (0.612, 0.662)
PGS*	0.709 (0.696, 0.722)	0.680 (0.663, 0.697)	0.705 (0.682, 0.728)
PXS*	0.762 (0.749, 0.775)	0.732 (0.716, 0.748)	0.774 (0.753, 0.795)
CRS*	0.839 (0.829, 0.849)	0.817 (0.804, 0.830)	0.855 (0.838, 0.872)
PGS+PXS*	0.776 (0.764, 0.788)	0.749 (0.734, 0.764)	0.786 (0.765, 0.807)
CRS+PGS*	0.844 (0.834, 0.854)	0.821 (0.808, 0.834)	0.859 (0.842, 0.876)
CRS+PXS*	0.850 (0.840, 0.860)	0.829 (0.816, 0.842)	0.866 (0.850, 0.882)
CRS+PXS+PGS*	0.855 (0.845, 0.865)	0.834 (0.821, 0.847)	0.869 (0.853, 0.885)

PRS: Khera et al, Nature Genetics 2018
PXS: 12 non-genetic factors (selected by *XWAS* plus LASSO)
CRS: FamHx, BP, BMI, glucose, HDL, triglycerides

Noble et al.: AUC 0.6-0.9 (BMJ, 2011)
Meigs et al.: C-index 0.9 (NEJM, 2008)

Diabetes Care 2021

A PXS may have utility those at highest aggregate risk or for reclassification of the CRS



		Hazard Ratio (95% CI)		
		PGS	**PXS**	**CRS**
Top % score	**1%**	2.64 (1.87, 3.73)	9.74 (7.96, 11.91)	15.11 (12.74, 17.92)
	5%	2.27 (1.90, 2.71)	6.72 (5.92, 7.63)	10.54 (9.39, 11.83)
	10%	2.00 (1.73, 2.31)	5.90 (5.28, 6.61)	9.97 (8.94, 11.13)
	20%	1.96 (1.75, 2.21)	4.72 (4.23, 5.27)	9.51 (8.44, 10.71)

Diabetes Care 2021

A PXS may have utility those at highest aggregate risk or for reclassification of the CRS

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	0.116 (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)

C

	CRS+PGS+PXS Model		
CRS Model	# Participants	Continuous NRI	Categorical NRI
Cases	1281	0.216 (0.182 to 0.275)	0.144 (0.105 to 0.194)
Noncases	67018	0.215 (0.186 to 0.238)	-0.011 (-0.016 to -0.007)
Full population	68299	0.431 (0.377 to 0.503)	0.132 (0.098 to 0.179)

(see also Elliott et al, JAMA 2020)

<u>Undiagnosed Diabetes (A1C > 6.5%)</u>

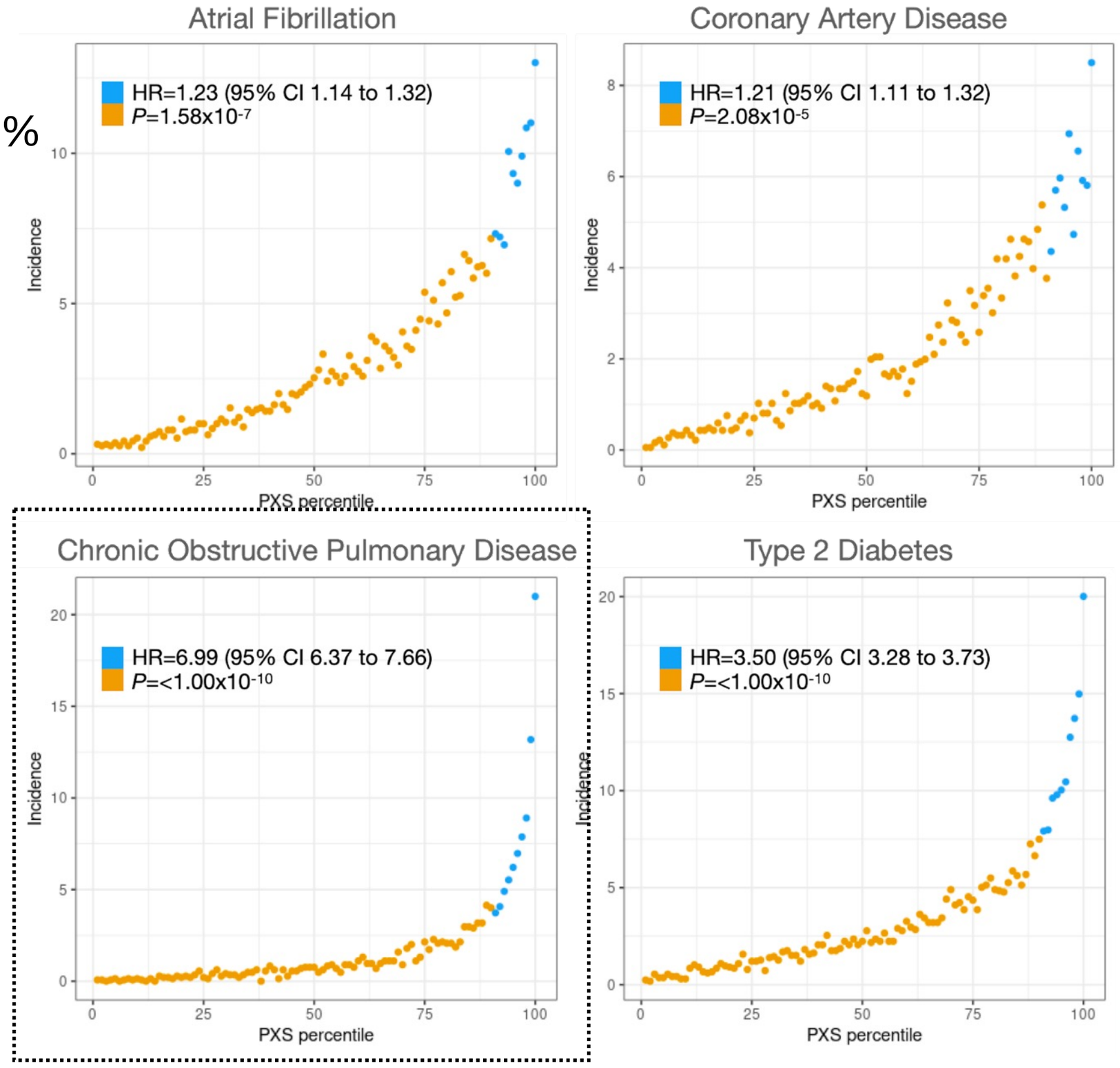
PRS: 0.696 (0.688, 0.705)

PXS: 0.756 (0.748, 0.764)

Diabetes Care 2021

PXStools: interrogating multiple disease outcomes demonstrates heterogeneity of predictions in UK Biobank

HR:
Top 10% vs. 90%

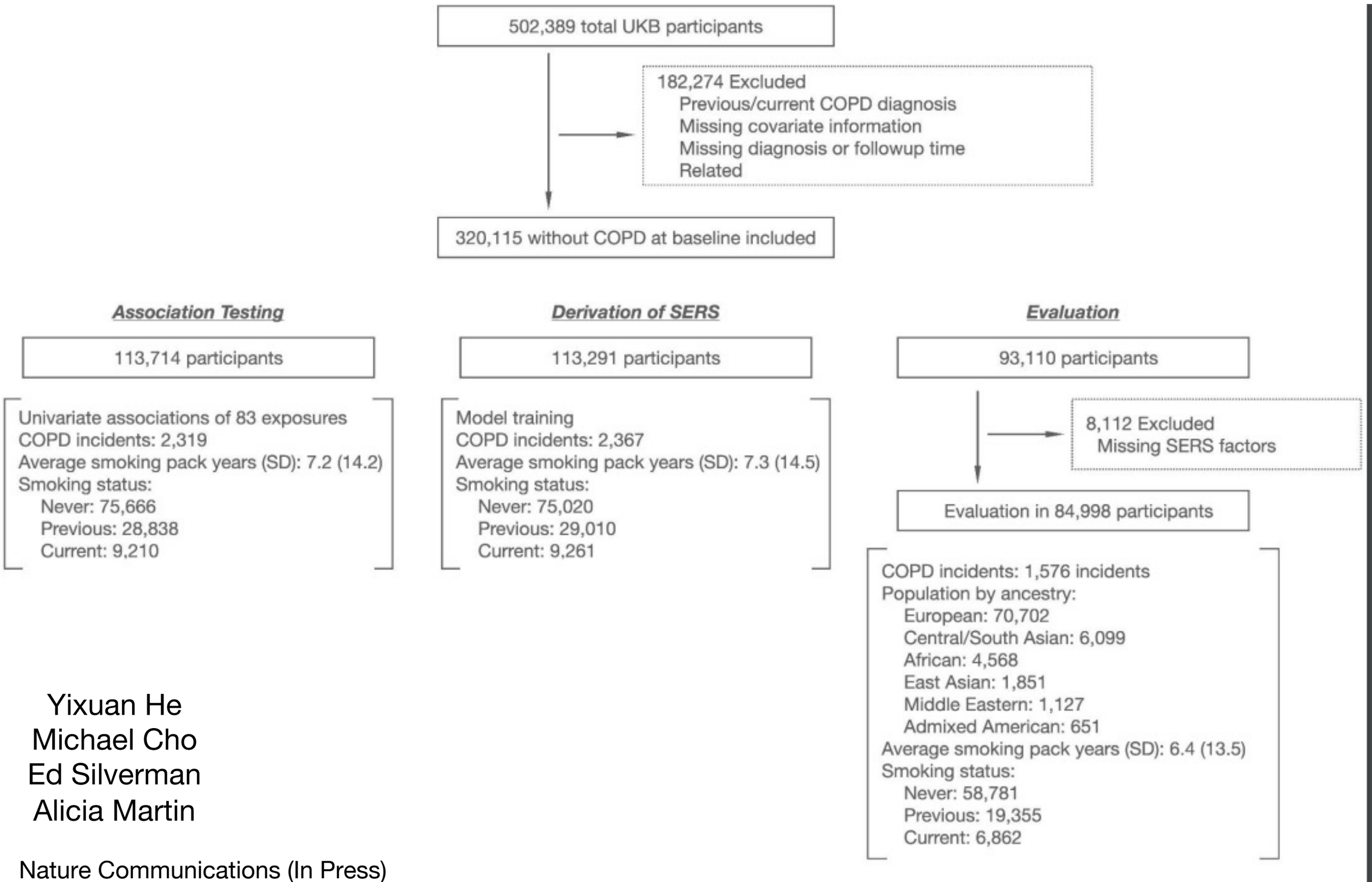


ExWAS
PXS via LASSO
Group LASSO

https://github.com/yixuanh/PXStools

Int J Epidemiology 2023

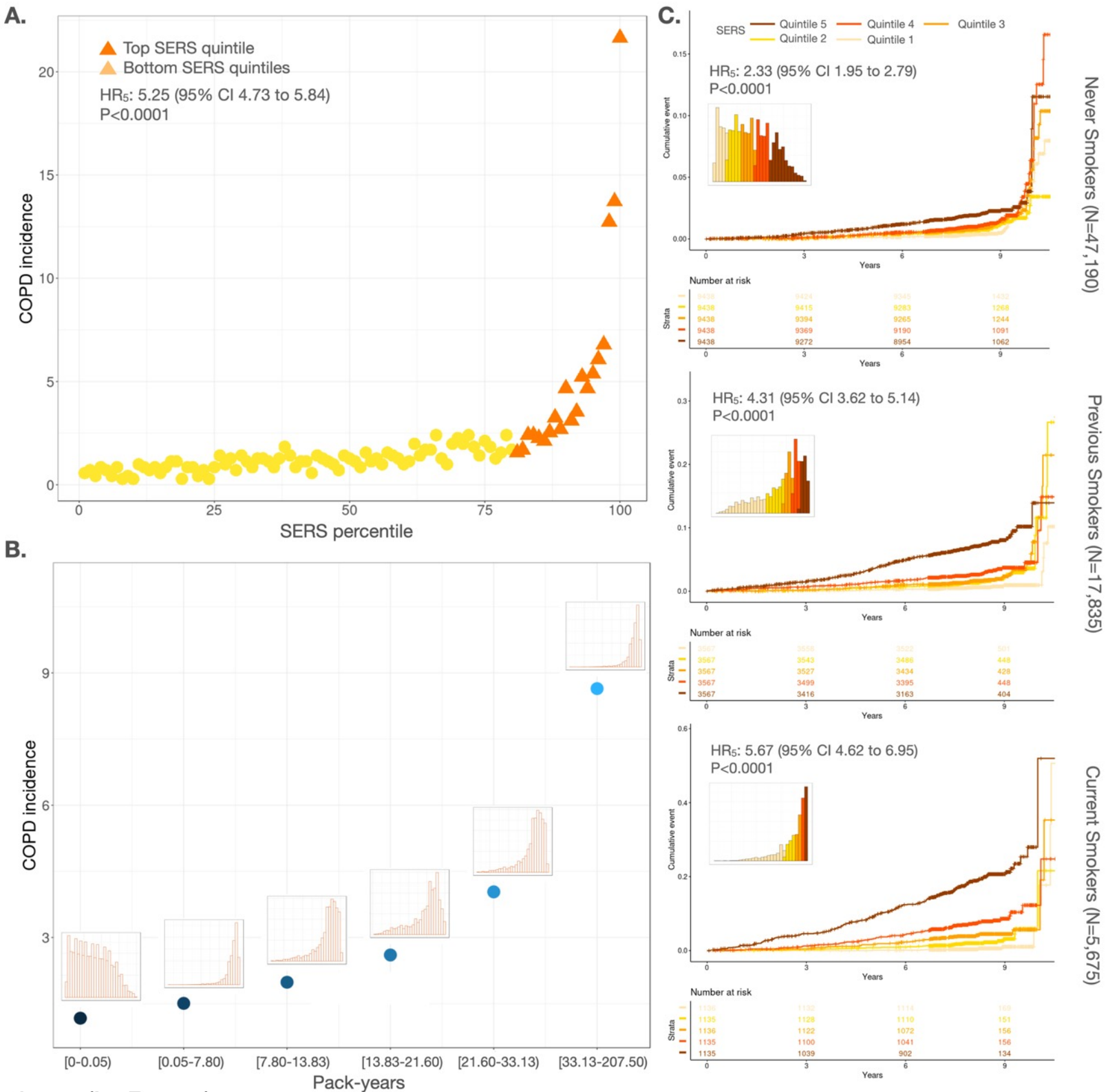
Building a socioeconomic and exposomic risk score for screening for COPD while considering smokers



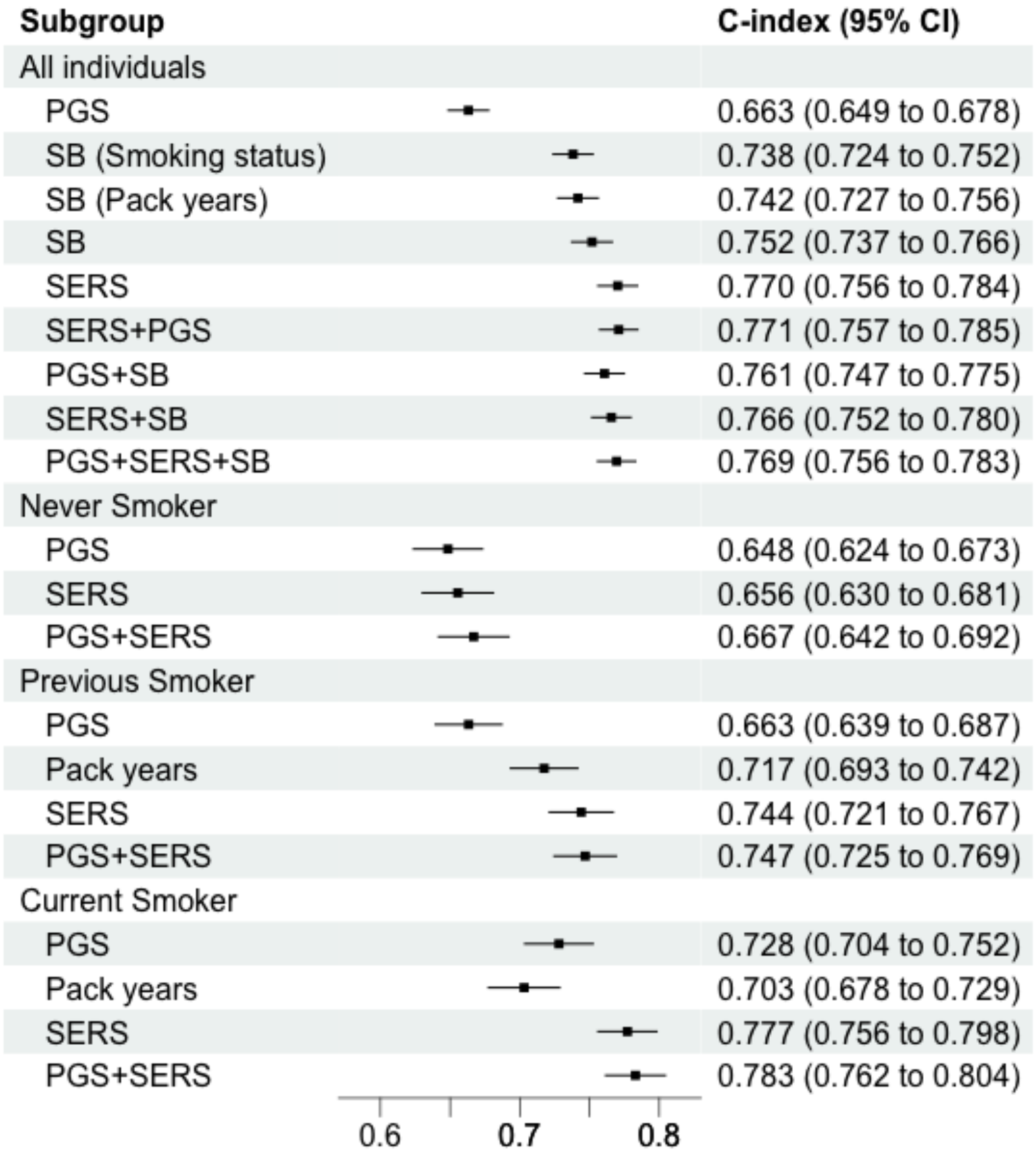
Yixuan He
Michael Cho
Ed Silverman
Alicia Martin

Nature Communications (In Press)

Building a socioeconomic and exposomic risk score for screening for COPD for smokers and non-smokers



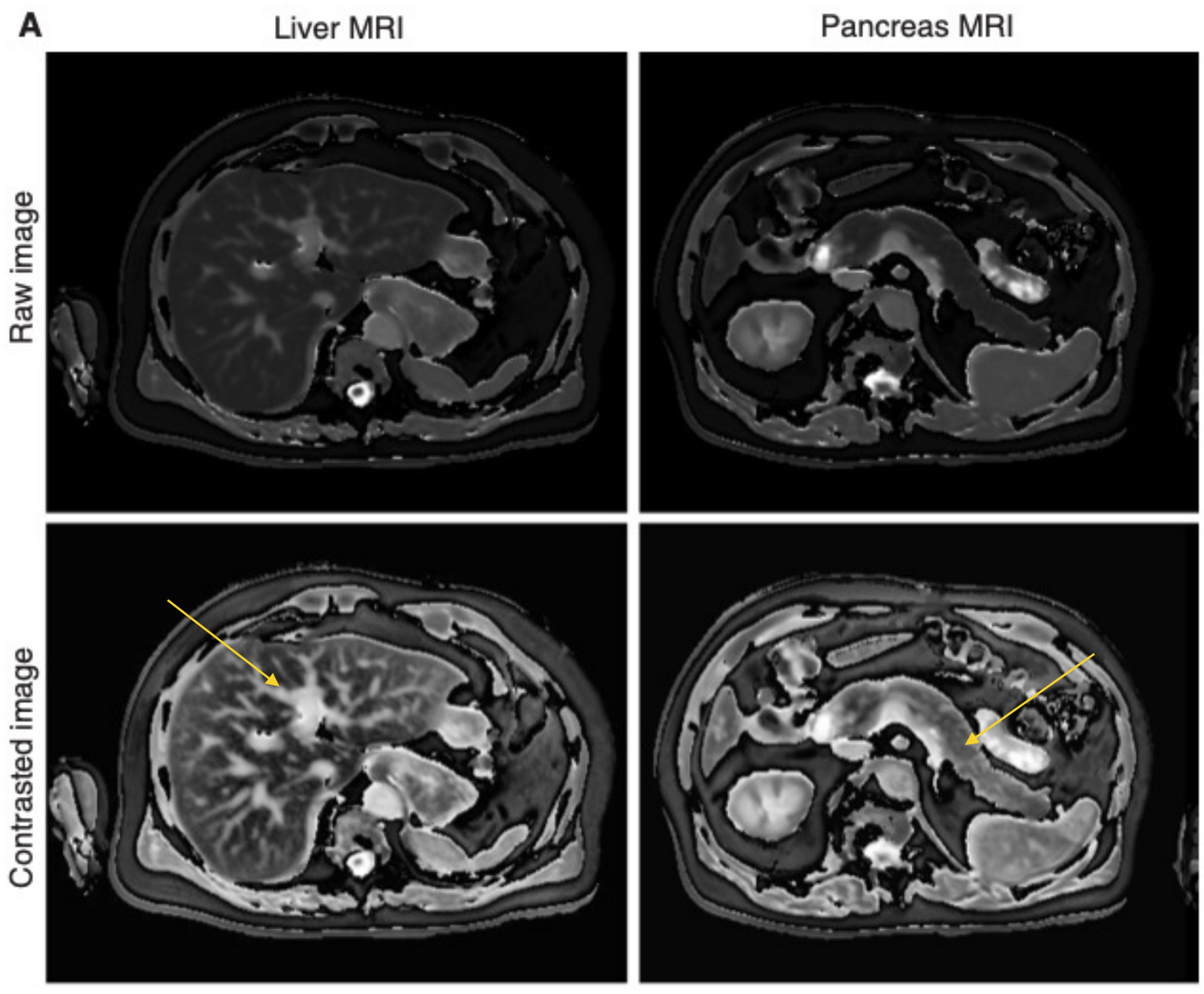
Nature Communications (In Press)



SERS provides improved prediction for smokers;
Comparable to genetics of lung function

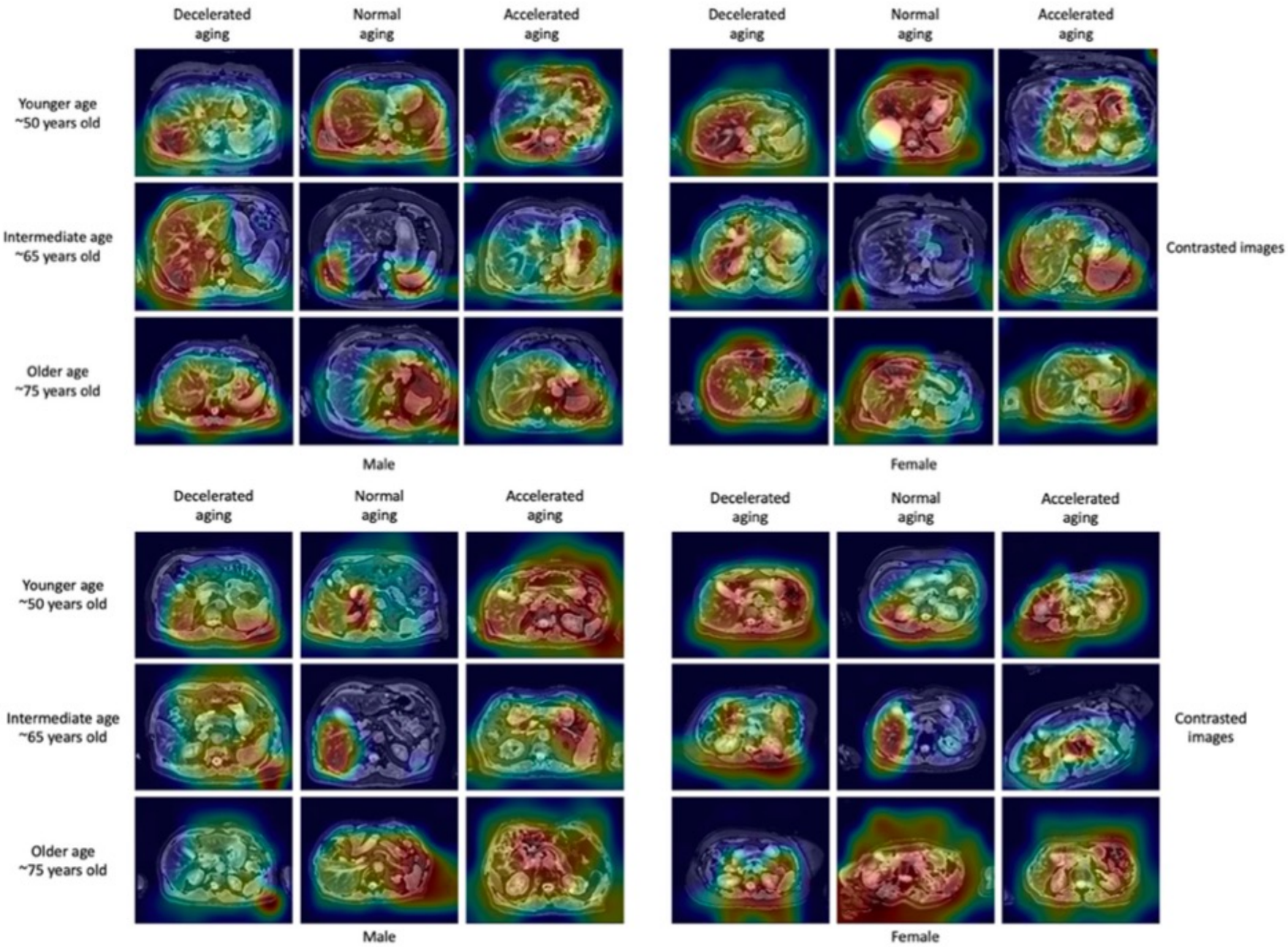
Nature Communications (In Press)

We predicted abdominal, pancreatic, and liver age with $R^2 > 70\%$ (MAE of 3.5 years) using convolutional neural networks (transfer learning)



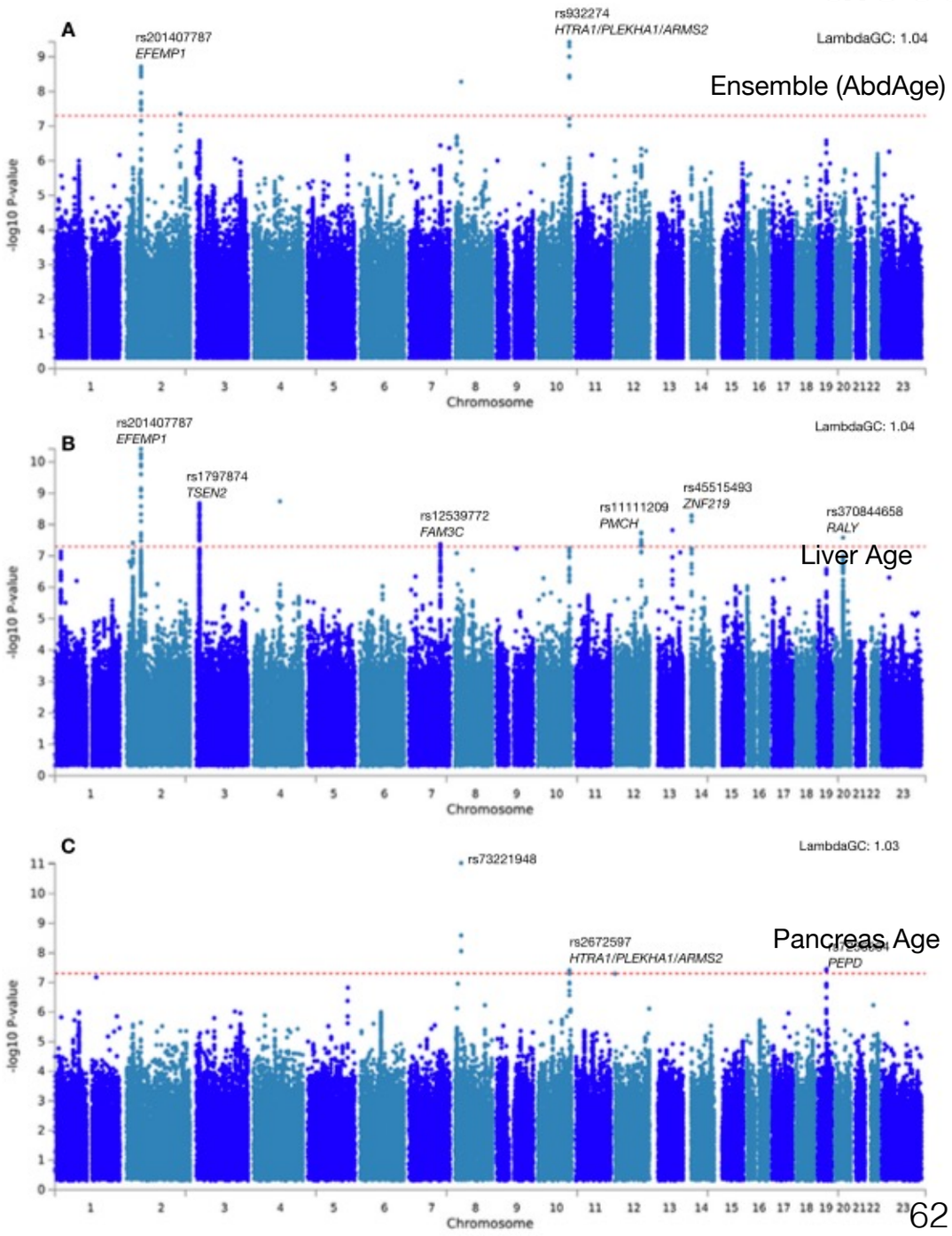
Le Goallec et al, *Nature Communications* 2022

Attention maps highlighted the liver, pancreas (but also the stomach, and surrounding adipose tissue)



Le Goallec et al, *Nature Communications* 2022

Abdominal, Pancreatic, Liver Age is heritable (h^2 of 22-26%), with GWAS signals implicated in metabolic disease



Genetic correlation between pancreas and liver: 0.86

Different GWAS hits for liver and pancreas dimensions suggest different aging processes

EFEMP1 (liver) is implicated in age-related macular degeneration

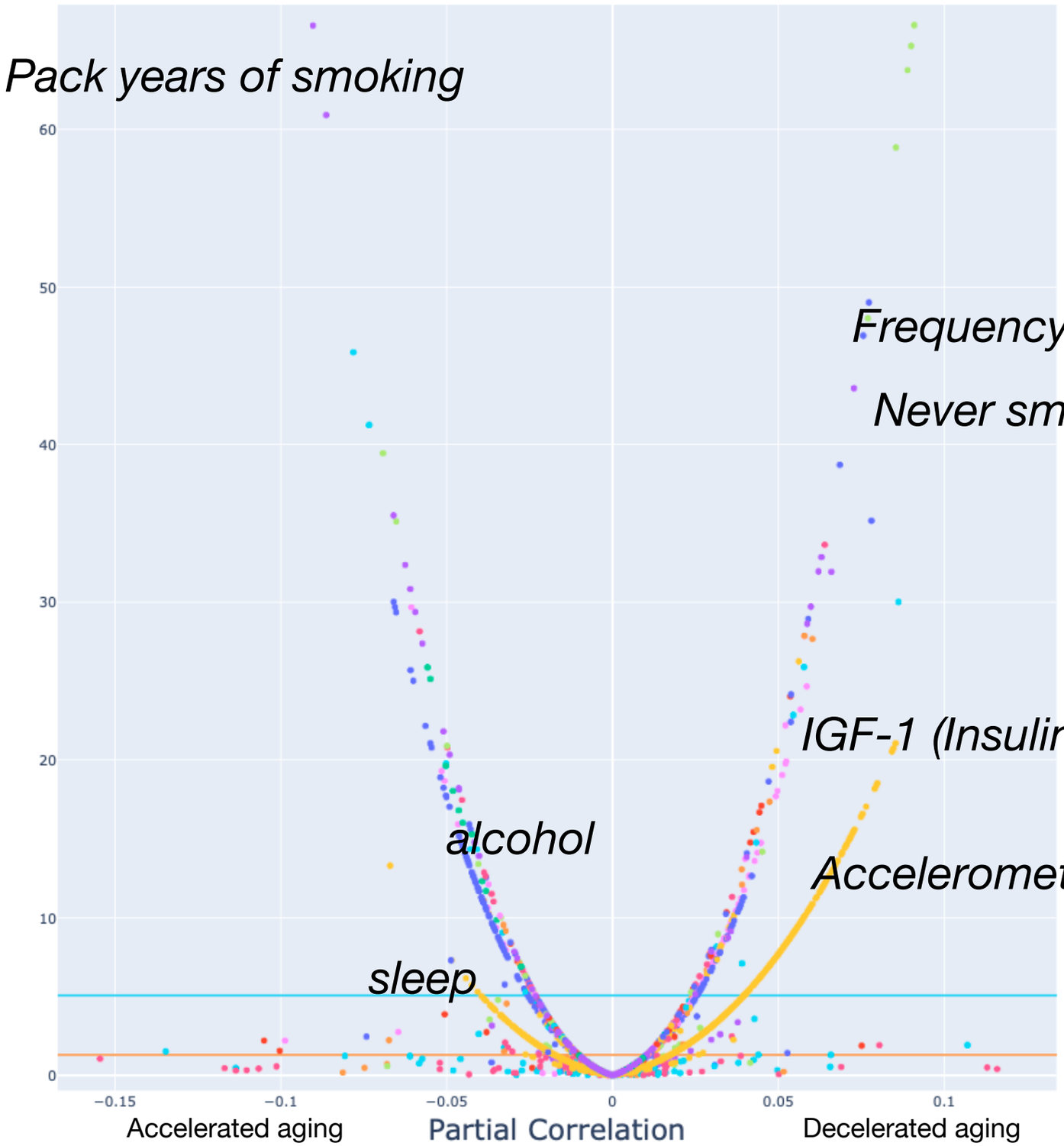
*PLEKHA1 (*pancreas) *shared* in type 2 diabetes, obesity

Genetic association distinct from T2D

Le Goallec et al, *Nature Communications* 2022

https://t2d.hugeamp.org/downloads.html

ExWAS (m=266) in abdominal aging: smoking, diet, physical activity, and alcohol (R^2 of ~2%)

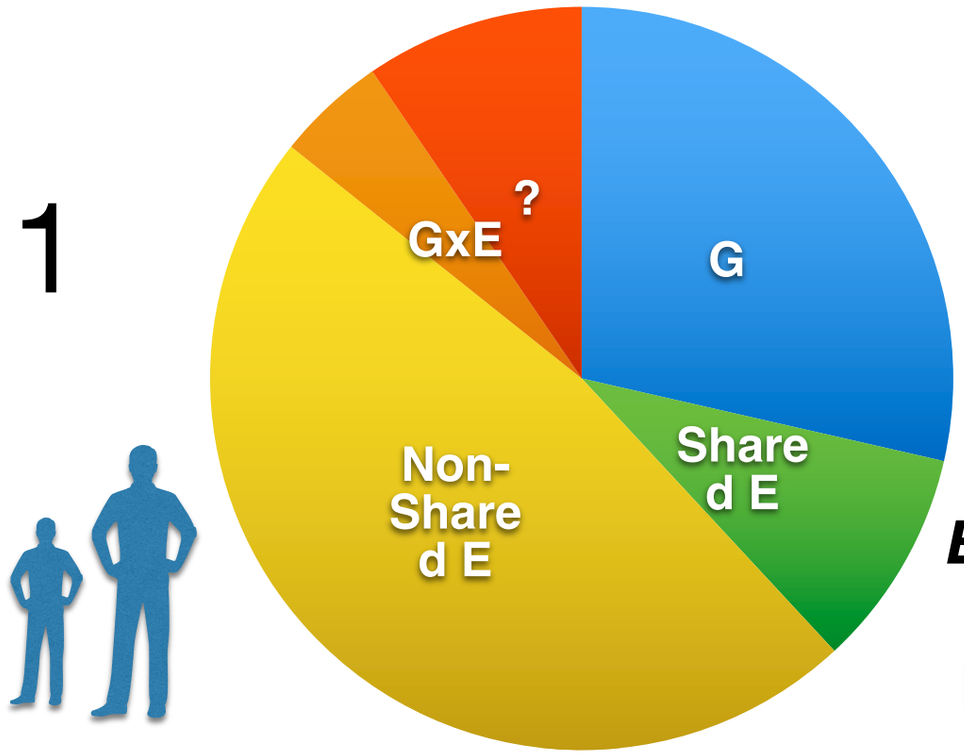


The lifestyle exposome intersects pancreas/abdominal aging and diabetes: What comes first - are they independent?

https://www.multidimensionality-of-aging.net/xwas

Key data applications for exposomic research:
(1) How much ***variation attributable to E*** in disease?
(2) What ***factors of the exposome*** are associated with disease?

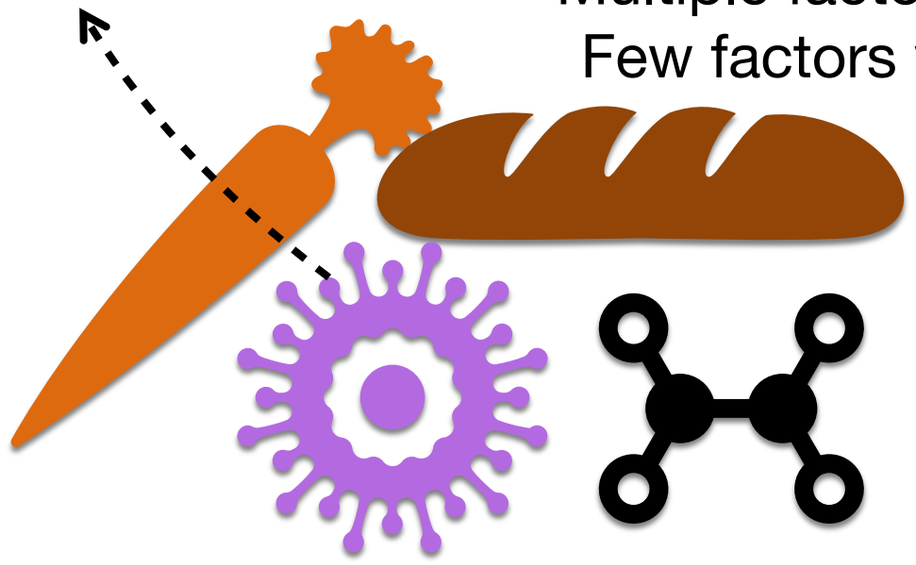
1



Larger the proportion (slice of pie):
More efficient discovery?

Exposure-wide association studies (ExWAS):
What factors are associated?
How do the exposures “add” up in aggregate?
Multiple factors of small effects?
Few factors with large effects?

2



?

Key applications of exposomic research: toward *ExWAS* and high-throughput epidemiology in biobanks

- Shared exposome explains 10% of total phenotype variation, and area-level socioeconomics explains 1%; where is the rest of the variation in most traits?
- New approaches to actualize the exposome to dissect social determinants from genetics and environment
- Big big data = big bias in non-genetic research, including identifying confounders to elucidate causality
- New ‘omics and imaging tools to examine the multidimensionality of disease, such as aging

Acknowledgements

RagGroup

Pedram Fard
Sivateja Tangirala
Yixuan He
Alan LeGoallec
Jake Chung
Chirag Lakhani
Vy Nguyen

Mentioned Collaborators

John Ioannidis
Peter Visscher
Arjun Manrai
Francesca Dominici
Alicia Martin
Hossein Estiri
Eran Bendavid

Funding

NIEHS R01 ES032470
NIA RF1 AG74372
NIAID R01 AI127250
NIDDK T32 110919

National Institute
of Allergy and
Infectious Diseases

HARVARD
MEDICAL SCHOOL

DEPARTMENT OF
Biomedical Informatics

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