Leveraging biobank scale to prioritize exposure-phenotype associations in children

Chirag J Patel

Children's Environmental Health: workshop on future priorities
National Academies of Sciences Engineering and Medicine
August 3, 2021

HARVARD
MEDICAL SCHOOL

DEPARTMENT OF
Biomedical Informatics

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

It is possible to identify new and established exposures associated with health in big ***biobanked*** data!

1. (1) Discover & replicate new exposures and genes;
2. (2) Interrogate new biological pathways;
3. (3) “Triangulate” possible causal relationships;
4. (4) Perform meta-analysis and synthesis

All of Us SM | The Precision Medicine Initiative®
THE FUTURE OF HEALTH BEGINS WITH YOU

lifelines

biobank uk

… what about to study early development and children?

Genomics and the *genome-wide association study:* an example of robust big data observational studies?!

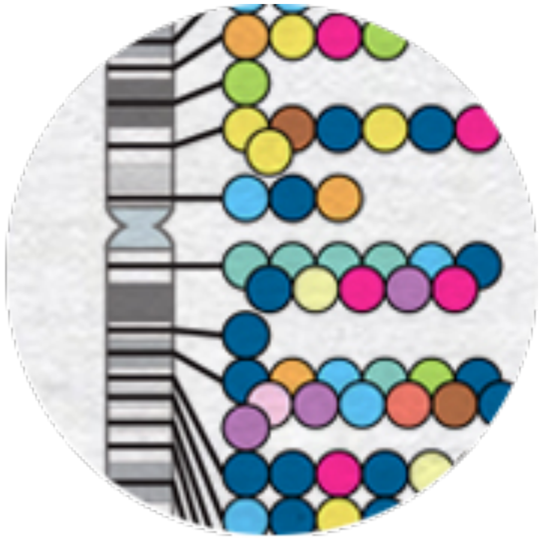
3,567 publications (as of 9/18/18)
71,673 ***G-P*** associations

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

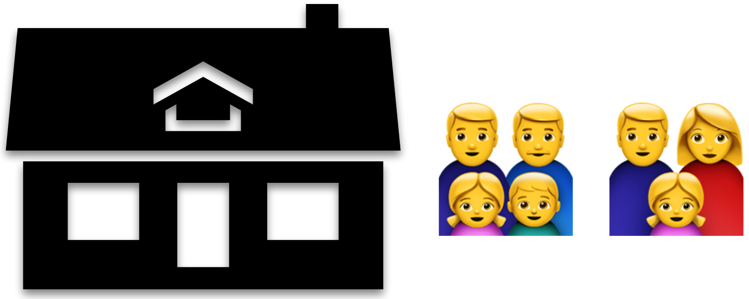
- Scaled for discovery
- Replicated associations
- Meta-analysis across cohorts
- New biological pathways
- Negligible confounding



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

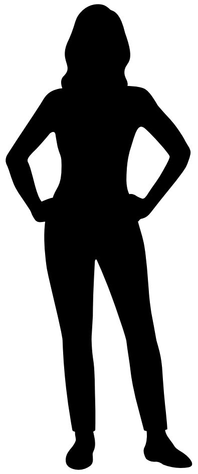
The exposome is ***shared*** and ***non-shared***!

shared



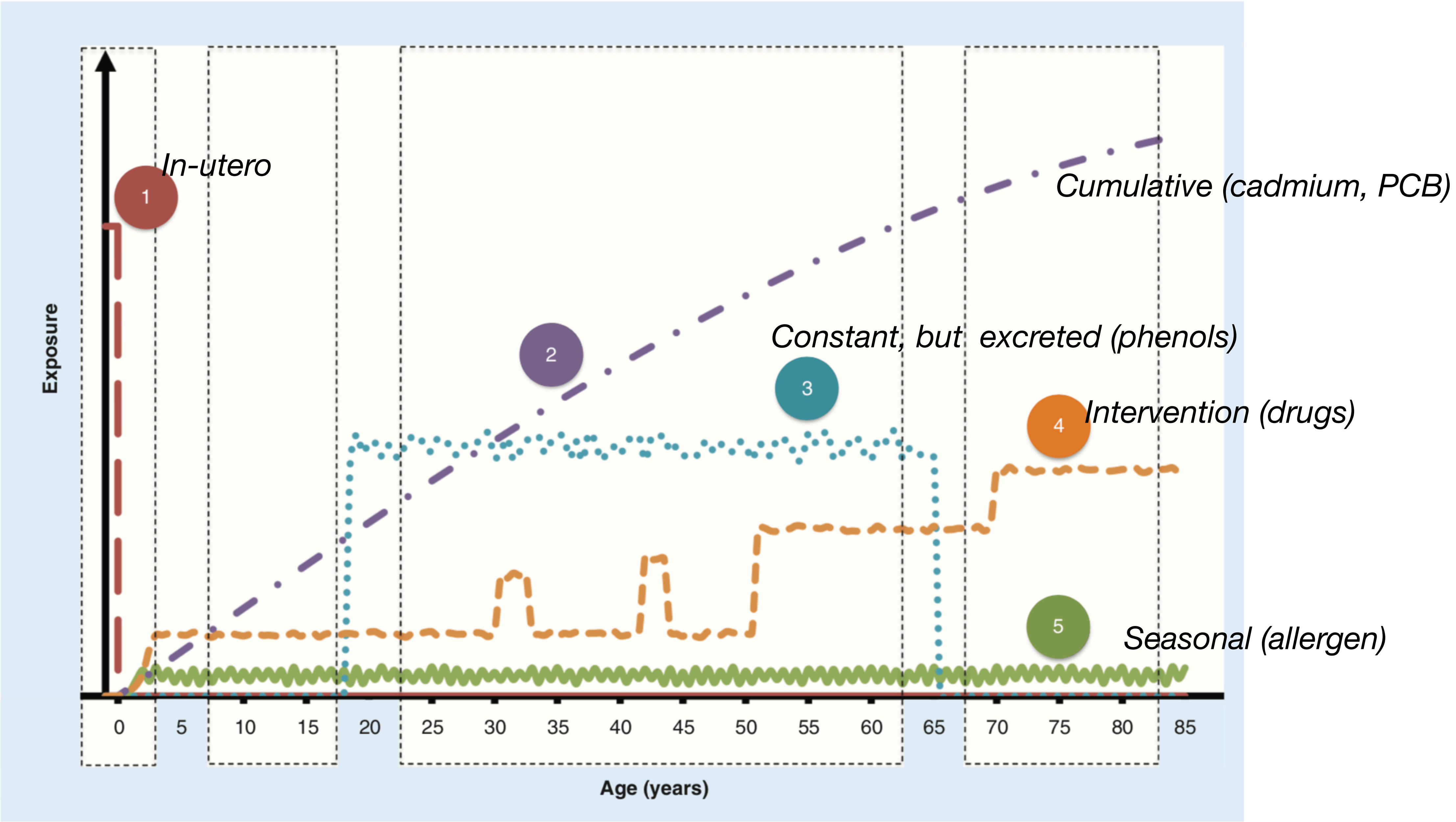
Small particles in air pollution
Extreme weather (heat and cold)
In-utero exposure

non-shared



Nutrients from dietary choice

Biobanking and capture of exposure at different times during development: *when (and at what frequency) do we measure?*

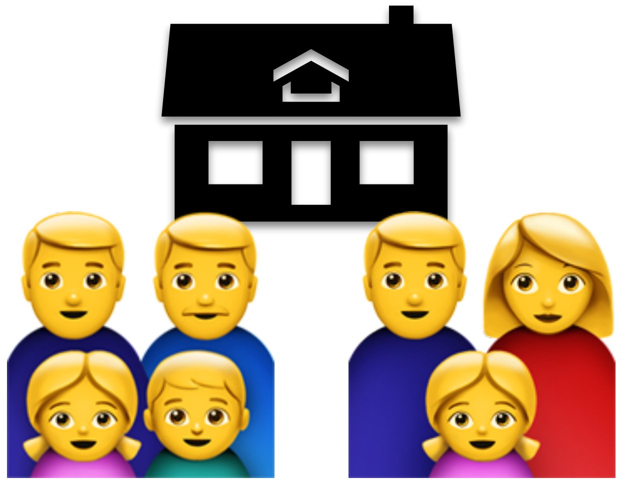


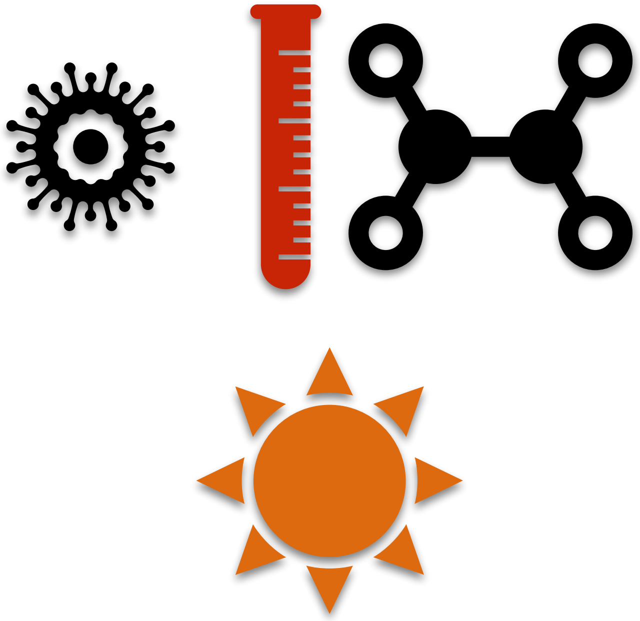
Not shown: Diurnal

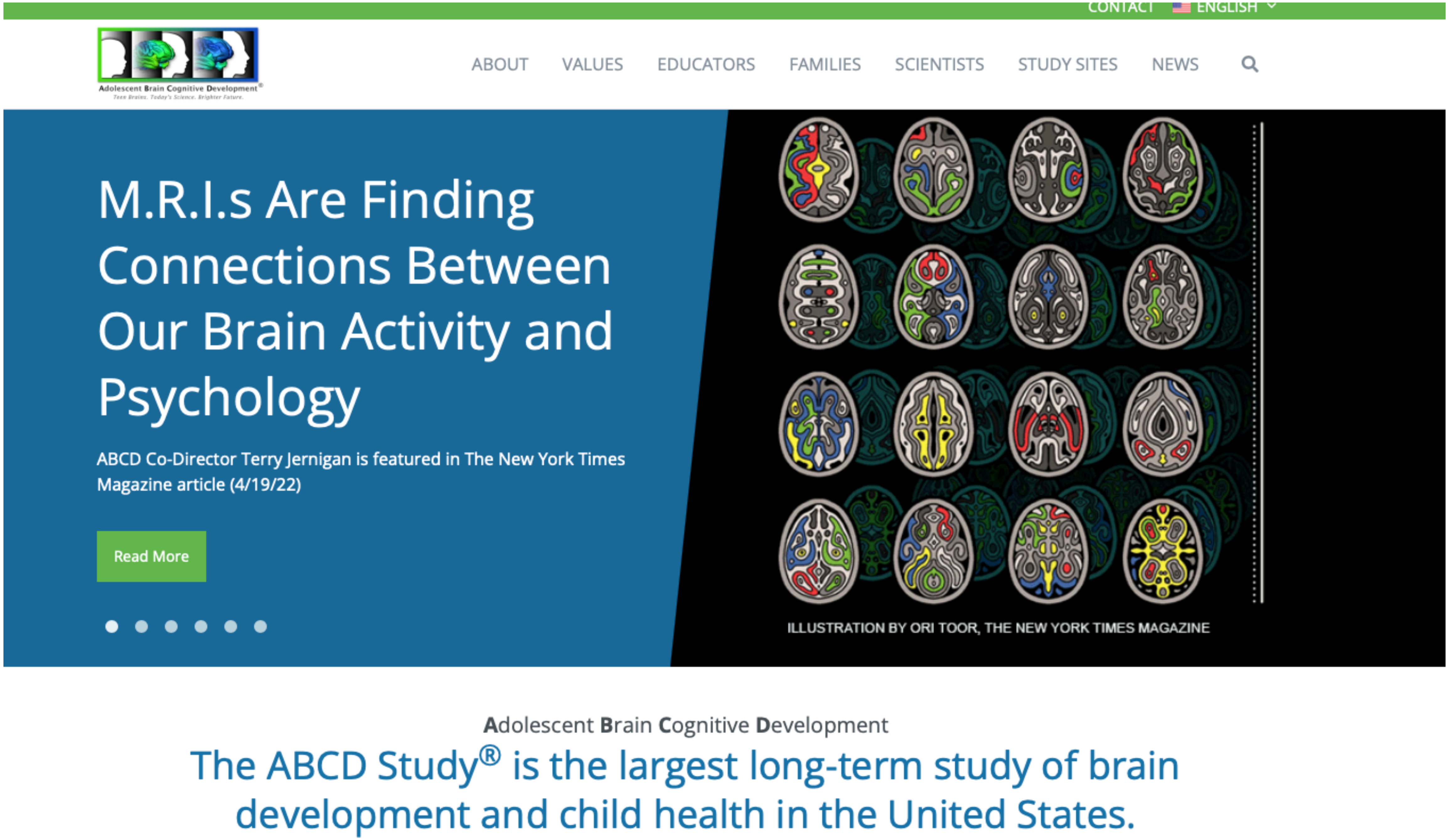
Athersuch *Bioanalysis* 2012

Requirements for biobanked study of children to identify exposures associated with health and disease

- Consent (Kilma et al, *Genetics in Medicine* 2014)
- Measurement of development-relevant phenotypes
- Frequent measure through *in-utero* and developmental time
- Biosamples to assay the ***exposome***
- Linkages to health information of mom & dad
- Associations with future health outcomes (adolescence, adulthood)
- ***Geospatial*** exposome biomarkers: climate and air pollution
- Data approaches to harmonize across cohorts for meta-analyses and systematic reviews






CONTACT
ENGLISH
Adolescent Brain Cognitive Development®
Teen Brains. Today's Science. Brighter Future.
ABOUT
VALUES
EDUCATORS
FAMILIES
SCIENTISTS
STUDY SITES
NEWS
Read More

M.R.I.s Are Finding Connections Between Our Brain Activity and Psychology

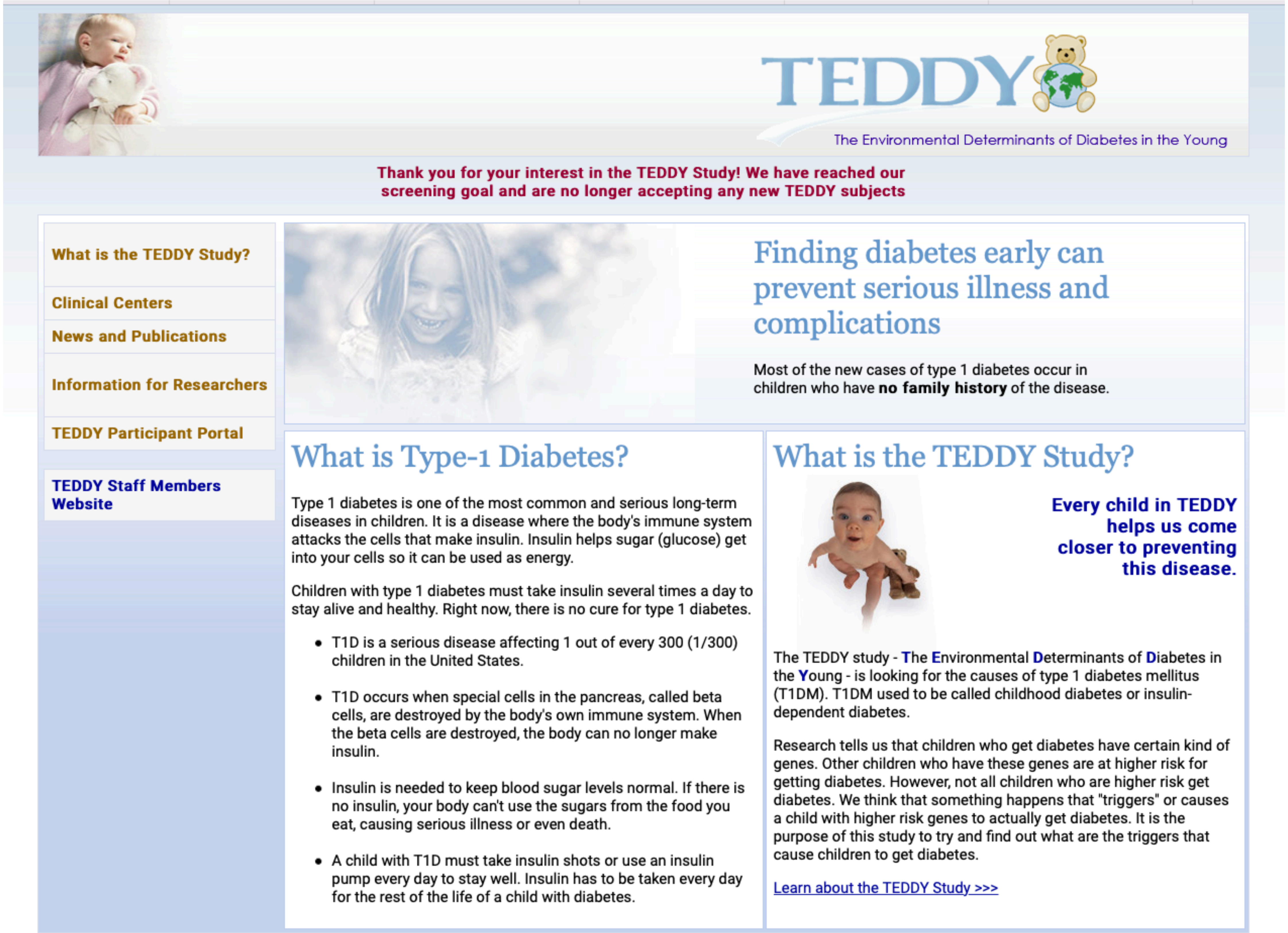
ABCD Co-Director Terry Jernigan is featured in The New York Times Magazine article (4/19/22)

ILLUSTRATION BY ORI TOOR, THE NEW YORK TIMES MAGAZINE

Adolescent **B**rain **C**ognitive **D**evelopment

The ABCD Study® is the largest long-term study of brain development and child health in the United States.

https://abcdstudy.org



TEDDY

The Environmental Determinants of Diabetes in the Young

Thank you for your interest in the TEDDY Study! We have reached our screening goal and are no longer accepting any new TEDDY subjects

- What is the TEDDY Study?
- Clinical Centers
- News and Publications
- Information for Researchers
- TEDDY Participant Portal
- TEDDY Staff Members Website

Finding diabetes early can prevent serious illness and complications

Most of the new cases of type 1 diabetes occur in children who have **no family history** of the disease.

What is Type-1 Diabetes?

Type 1 diabetes is one of the most common and serious long-term diseases in children. It is a disease where the body's immune system attacks the cells that make insulin. Insulin helps sugar (glucose) get into your cells so it can be used as energy.

Children with type 1 diabetes must take insulin several times a day to stay alive and healthy. Right now, there is no cure for type 1 diabetes.

- T1D is a serious disease affecting 1 out of every 300 (1/300) children in the United States.
- T1D occurs when special cells in the pancreas, called beta cells, are destroyed by the body's own immune system. When the beta cells are destroyed, the body can no longer make insulin.
- Insulin is needed to keep blood sugar levels normal. If there is no insulin, your body can't use the sugars from the food you eat, causing serious illness or even death.
- A child with T1D must take insulin shots or use an insulin pump every day to stay well. Insulin has to be taken every day for the rest of the life of a child with diabetes.

What is the TEDDY Study?

Every child in TEDDY helps us come closer to preventing this disease.

The TEDDY study - **T**he **E**nvironmental **D**eterminants of **D**iabetes in the **Y**oung - is looking for the causes of type 1 diabetes mellitus (T1DM). T1DM used to be called childhood diabetes or insulin-dependent diabetes.

Research tells us that children who get diabetes have certain kind of genes. Other children who have these genes are at higher risk for getting diabetes. However, not all children who are higher risk get diabetes. We think that something happens that "triggers" or causes a child with higher risk genes to actually get diabetes. It is the purpose of this study to try and find out what are the triggers that cause children to get diabetes.

Learn about the TEDDY Study >>>

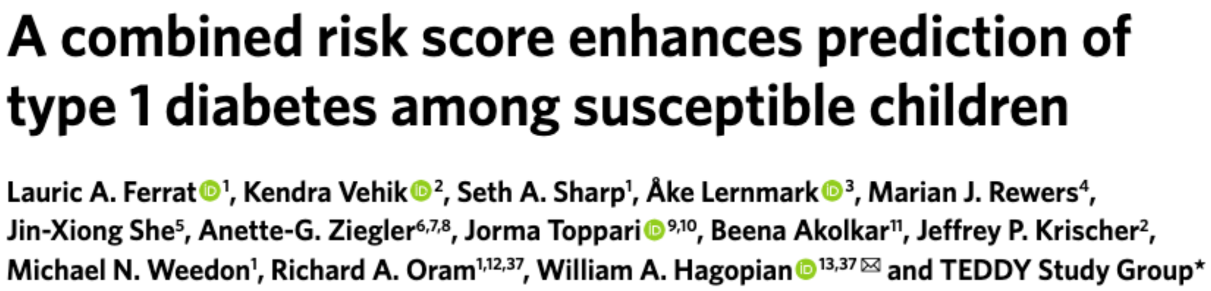
Family history (parents), genetic risk:
What triggers the onset?
https://teddy.epi.usf.edu/TEDDY/

Table 2. Follow-up Schedule

	Age in Months																								
	Screening		Follow-Up																						
Sampling Frequency	Birth	<4	2		3	4	5	6	7	8	9	10	11	12	15	18	21	24	27	12-48 mo Monthly Test	24-48 mo Every 3 mo Tests	24-48 mo Every 6 mo Tests	>48 mo Every 3 mo Tests	>48 mo Every 6 mo Tests	>48 mo Annual Tests
			Inform Parents of child's HLA risk	Mail initial enrollment and questionnaire packet																					
Blood**	X*	X*			X+			X+			X+			X+	X+	X+	X+	X+	X+		X+#			X+#	
Stool					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			X (until 10 years)	X (at 10 years); Collection stopped August 2018	
Tap Water											X									Collected every 2 years beginning at the 36 month visit					
Toenail Clippings																		X		Collected every 2 years beginning at the 24 month visit; Starting May 2017 collected every 1 year					
Salivary Cortisol																				Collected when child is 3.5, 4.5 and 5.5 years of age					
Nasal Swab											X			X	X	X	X	X	X		X#			X#	
Urine																						X (begins at 3 years)		X	
Primary Tooth	Collect when tooth naturally falls out - ages will vary																								
Weight and Length/Height Measurements; Body composition on some subjects					X			X			X			X	X	X	X	X	X		X#			X#	
Diet Questionnaires																									
-maternal pregnancy diet				X																					
-3 day diet record					X			X			X			X		X		X				X		X^	
Environmental Exposure Questionnaires																									
-maternal pregnancy/birth questionnaire				X																					
- parent questionnaire				X				X							X				X				Annually after 27 mos		
- child questionnaire																									X (begins at 10 years)
Demographic/Family History/Other questionnaire											X									Demographic data will be updated every 2 years thereafter; Family History data will be updated every 4 years thereafter					
TEDDY Book Extraction					X			X			X			X	X	X	X	X	X		X#			X#	
Child Behavior Checklist/ Strengths and Difficulties Questionnaire																					CBCL completed when child is 3.5, 4.5 and 5.5 years of age; SDQ completed by both parent and child when child is 11.5 and 13.5 years of age				
Physical Activity Assessment																									X (begins at 5 years)%
Pubertal Status Assessment																								X (begins at 8 years)	

*If cord blood is not available at birth for HLA typing then capillary blood should be drawn.

Risk score prediction of type 1 diabetes across age: do exposures interact through development for children at risk?



A combined risk score enhances prediction of type 1 diabetes among susceptible children

Lauric A. Ferrat[1], Kendra Vehik[2], Seth A. Sharp[1], Åke Lernmark[3], Marian J. Rewers[4], Jin-Xiong She[5], Anette-G. Ziegler[6,7,8], Jorma Toppari[9,10], Beena Akolkar[11], Jeffrey P. Krischer[2], Michael N. Weedon[1], Richard A. Oram[1,12,37], William A. Hagopian[13,37] ✉ and TEDDY Study Group*

Nature Medicine 2020

2 years of age

4 years of age

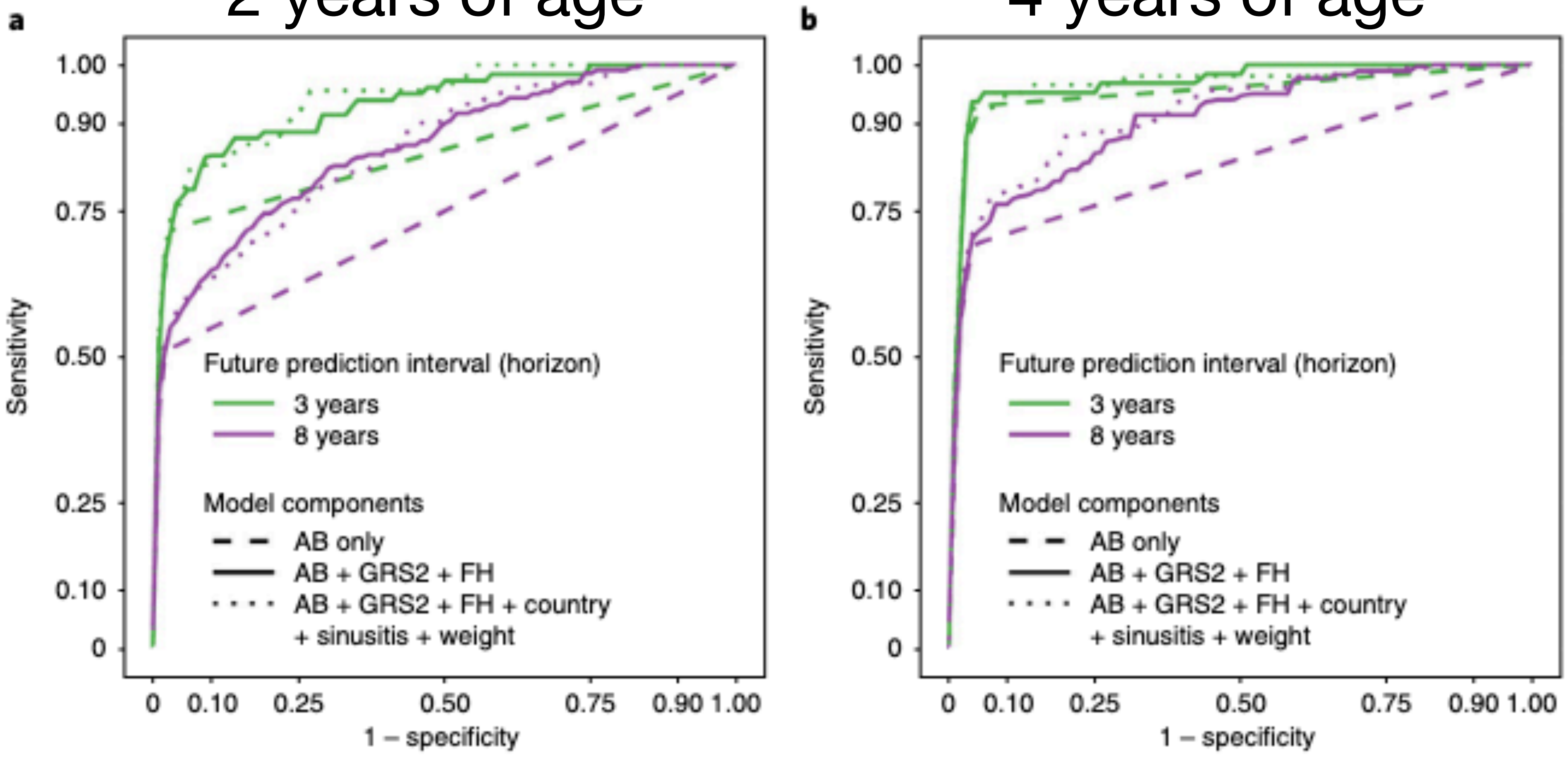


Fig. 2 | ROC curves derived from models incorporating different numbers of variables. a,b, Dotted, solid and dashed lines denote the use of all six variables, three variables and autoantibodies only, respectively. The curves in **a** use a landmark age of 2 years, with prediction horizons of 3 or 8 years, as indicated. The curves in **b** use a landmark age of 4 years, also with prediction horizons of 3 or 8 years, as indicated. AB, autoantibodies; FH, family history.

Health insurance claims data to partition the **genome** and **exposome** of phenotypes (<25 years of age)

Weather

Air Pollution

Census SES

+

insurance claims

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

N ~ 45M

Jian Yang
Peter M Visscher

Chirag Lakhani

Braden Tierney

Arjun Manrai

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!

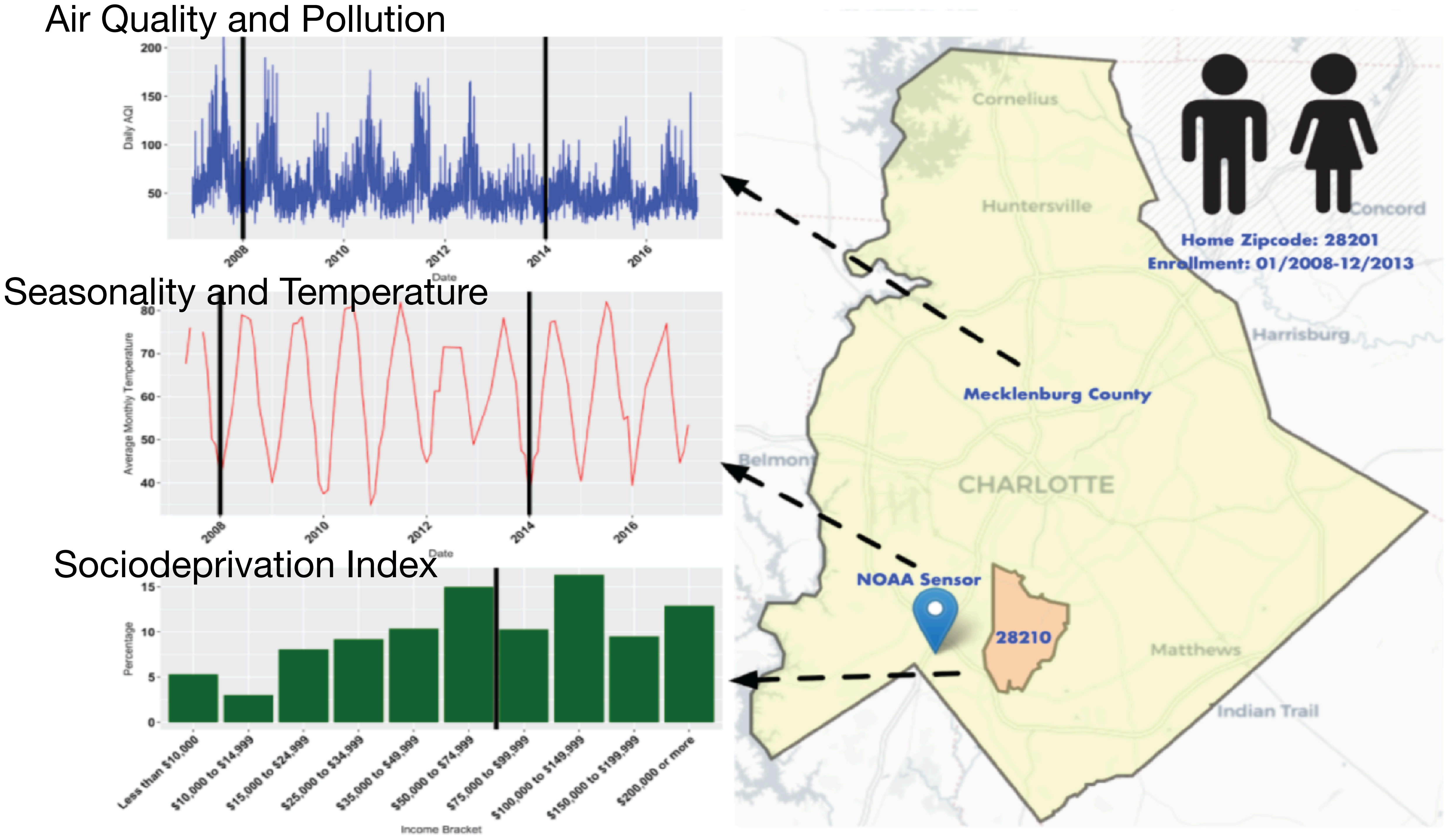
Largest collection of twins in US (next largest has ~28k pairs)

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***

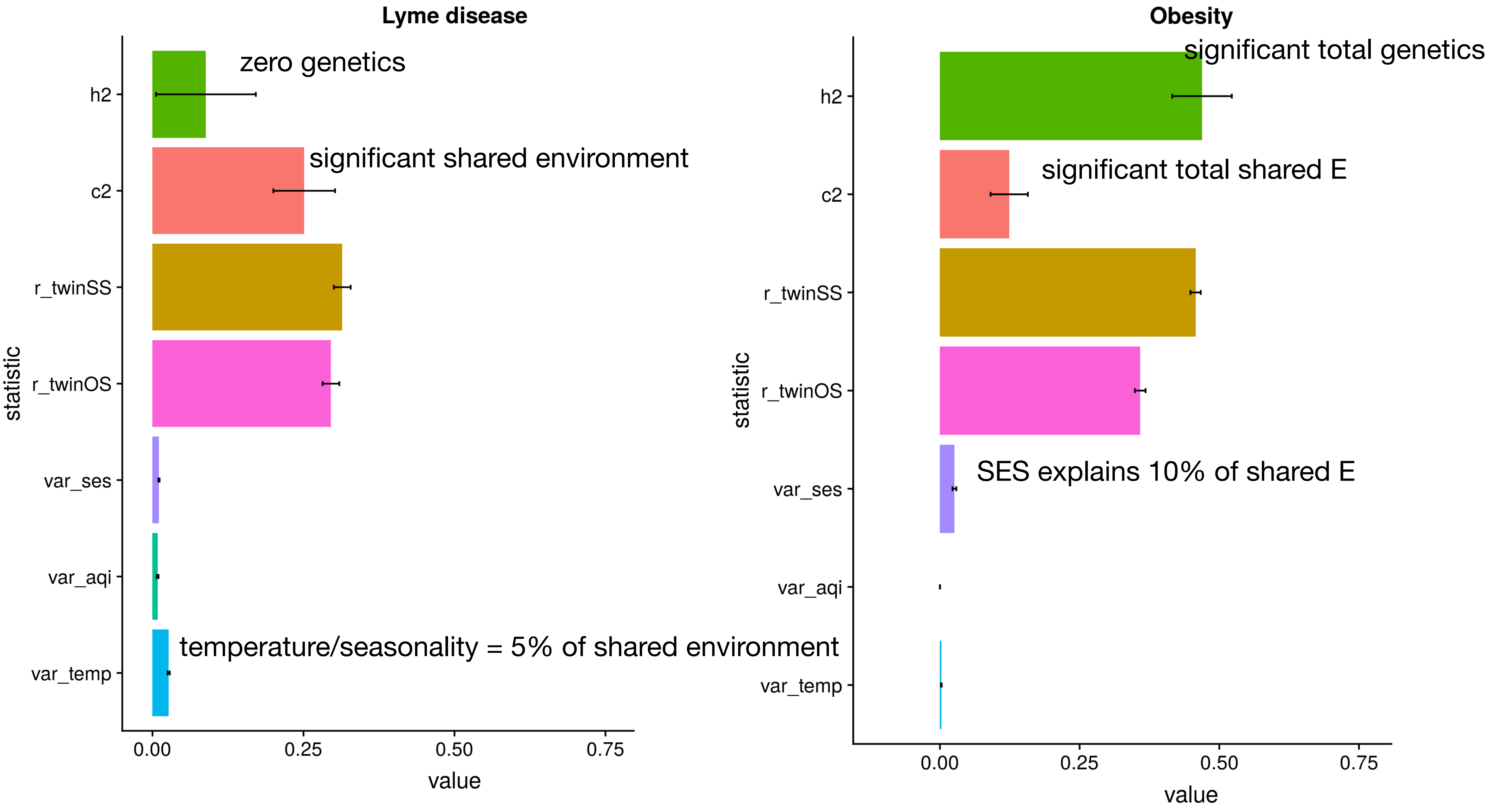


Air Quality and Pollution

Seasonality and Temperature

Sociodeprivation Index

Decomposing ***G*** (h^2), shared ***E*** (c^2) and factors of shared ***E*** (SES, air quality, and temperature) in 2 candidate phenotypes: Lyme and Obesity



zero genetics

significant shared environment

temperature/seasonality = 5% of shared environment

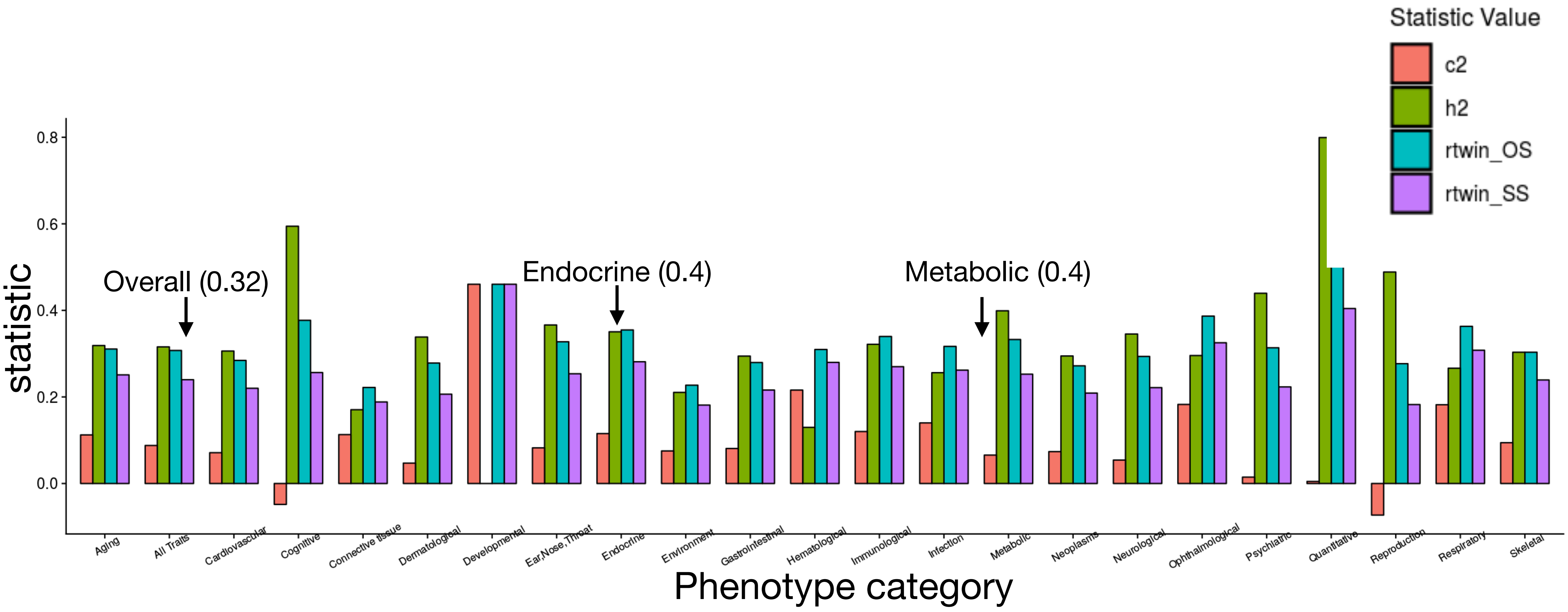
significant total genetics

significant total shared E

SES explains 10% of shared E

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***



Overall (0.32)

Endocrine (0.4)

Metabolic (0.4)

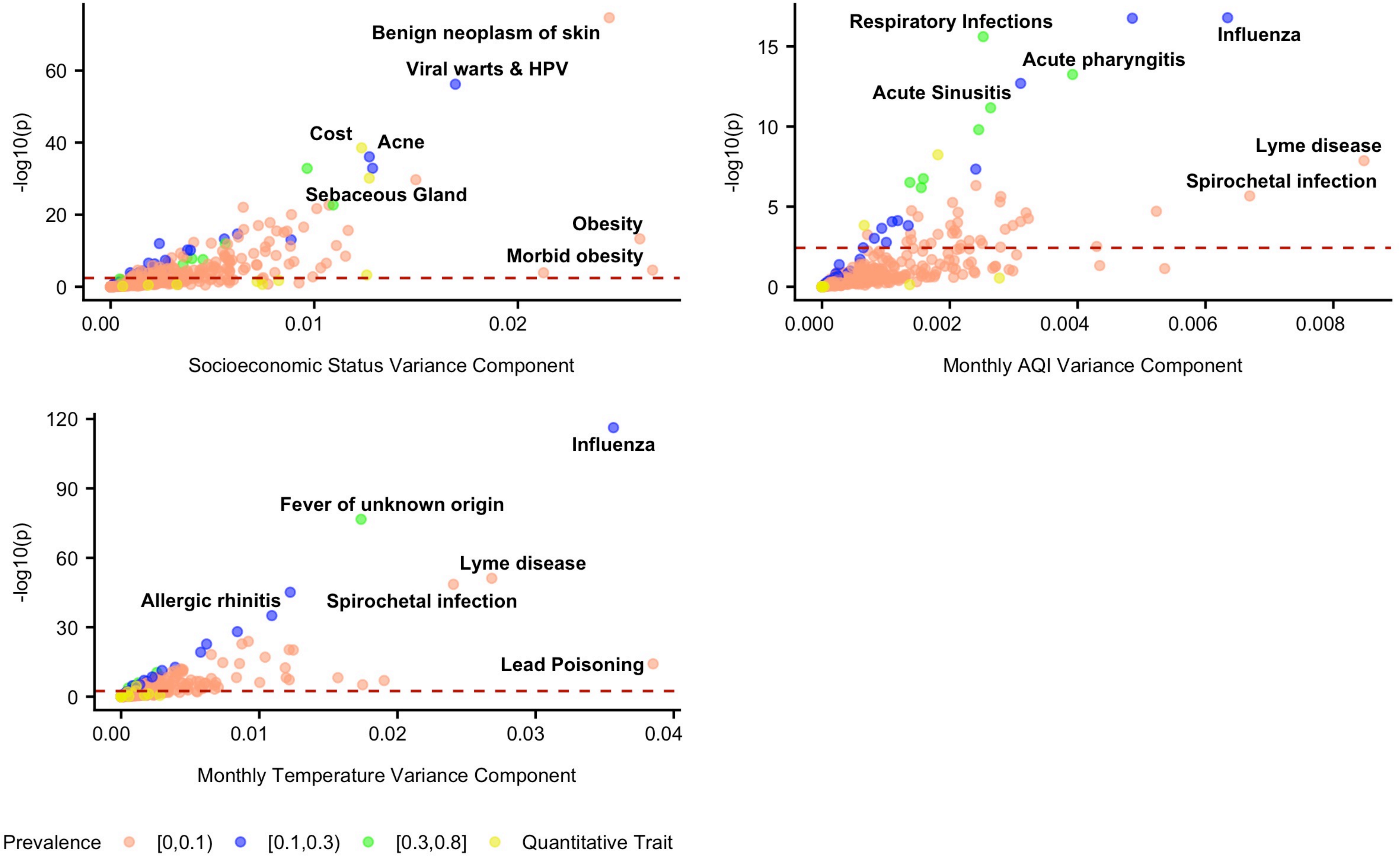
US-based, ages < 25

CaTCH: Claims analysis of Twin Correlation and Heritability

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

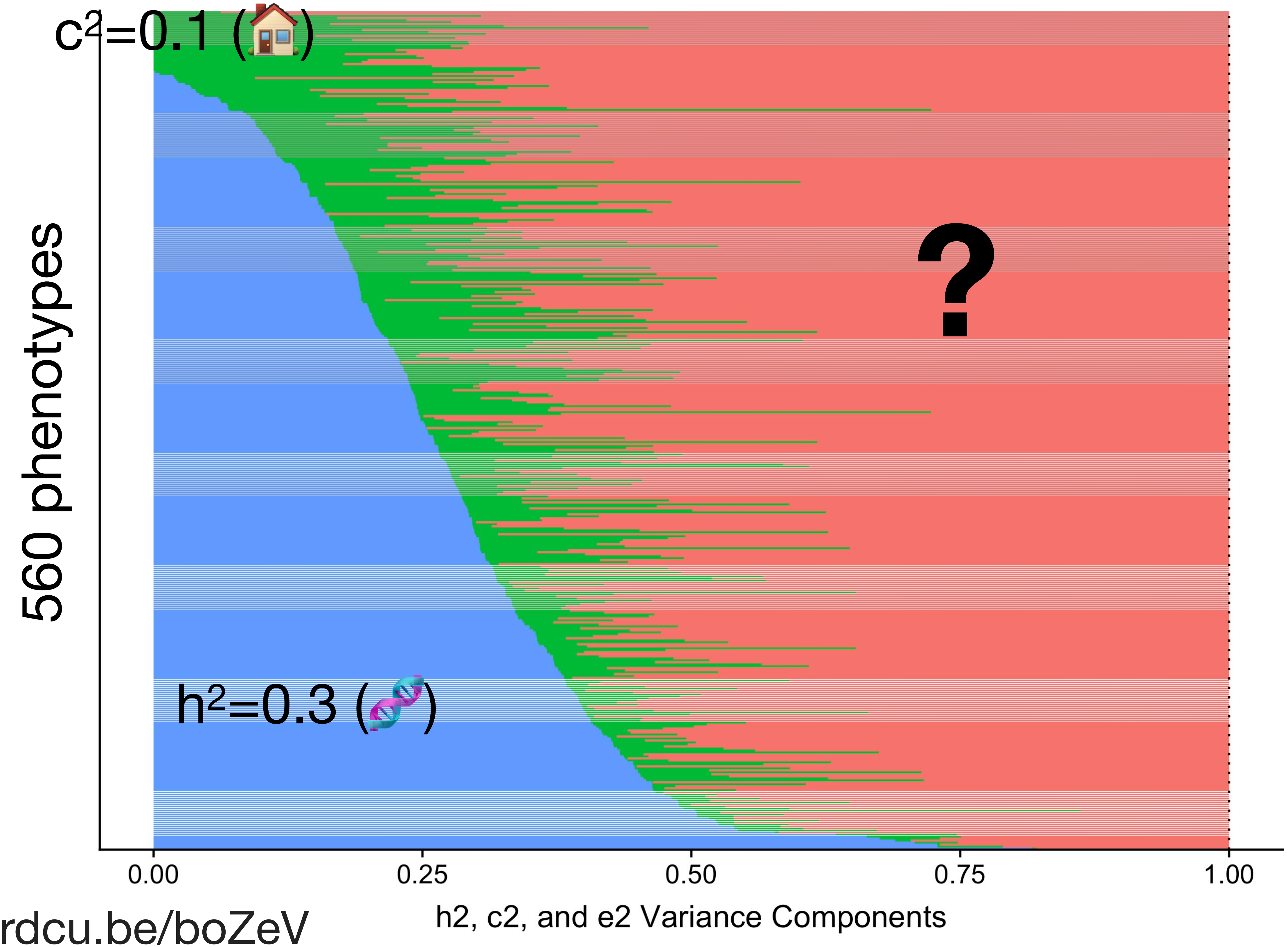
Lakhani et al., Nature Genetics 2019

Dissecting the role of air pollution, climate, and geocoded SES total shared environment



Lakhani et al., Nature Genetics 2019

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

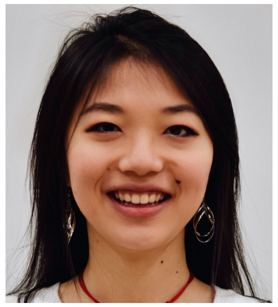


https://rdcu.be/boZeV

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

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

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***):
UK Biobank, 111 modifiable/non-modifable 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-modifable exposures

N=111

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

Filter & Select

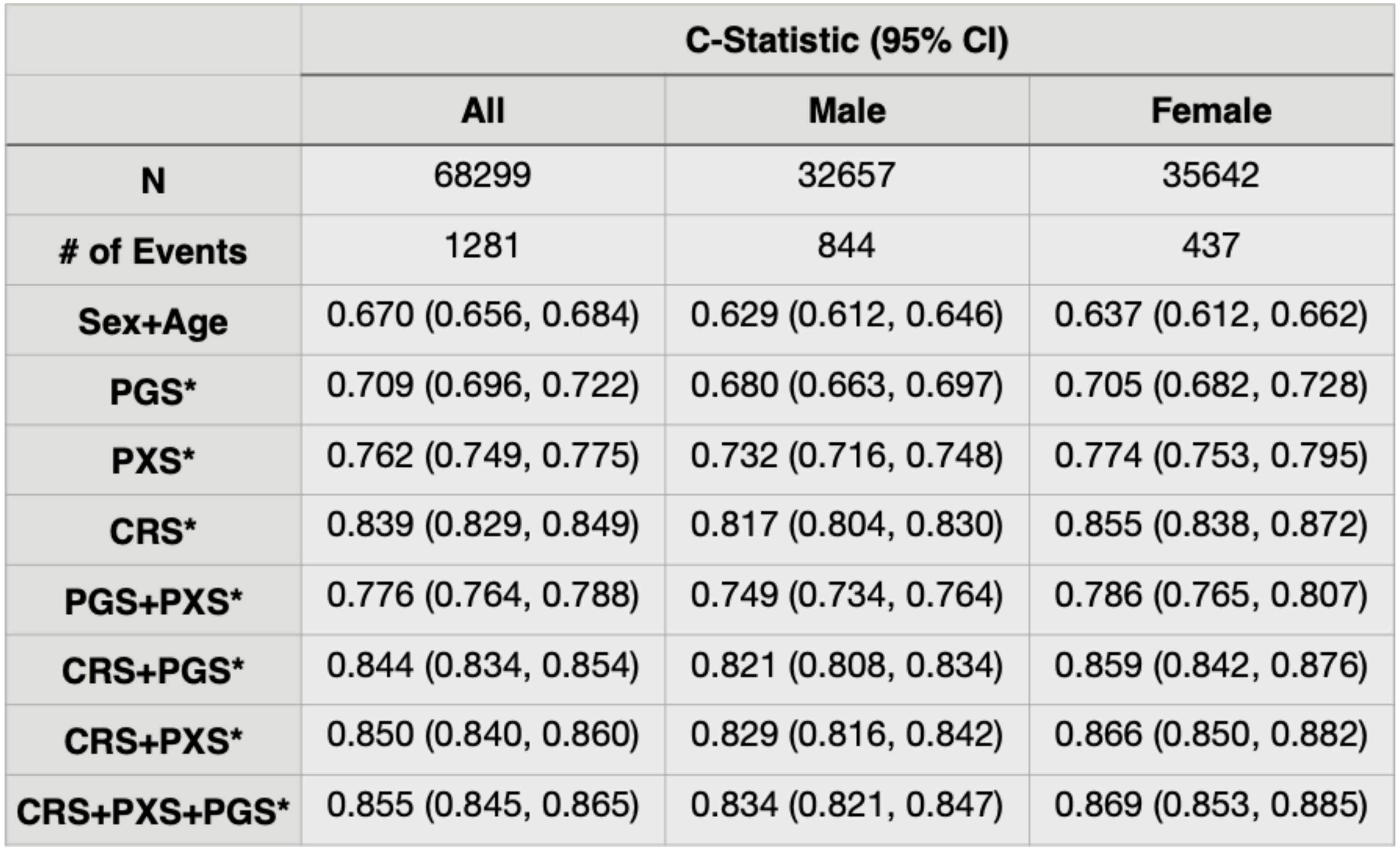
XWAS
Lasso
P value thresholds

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)



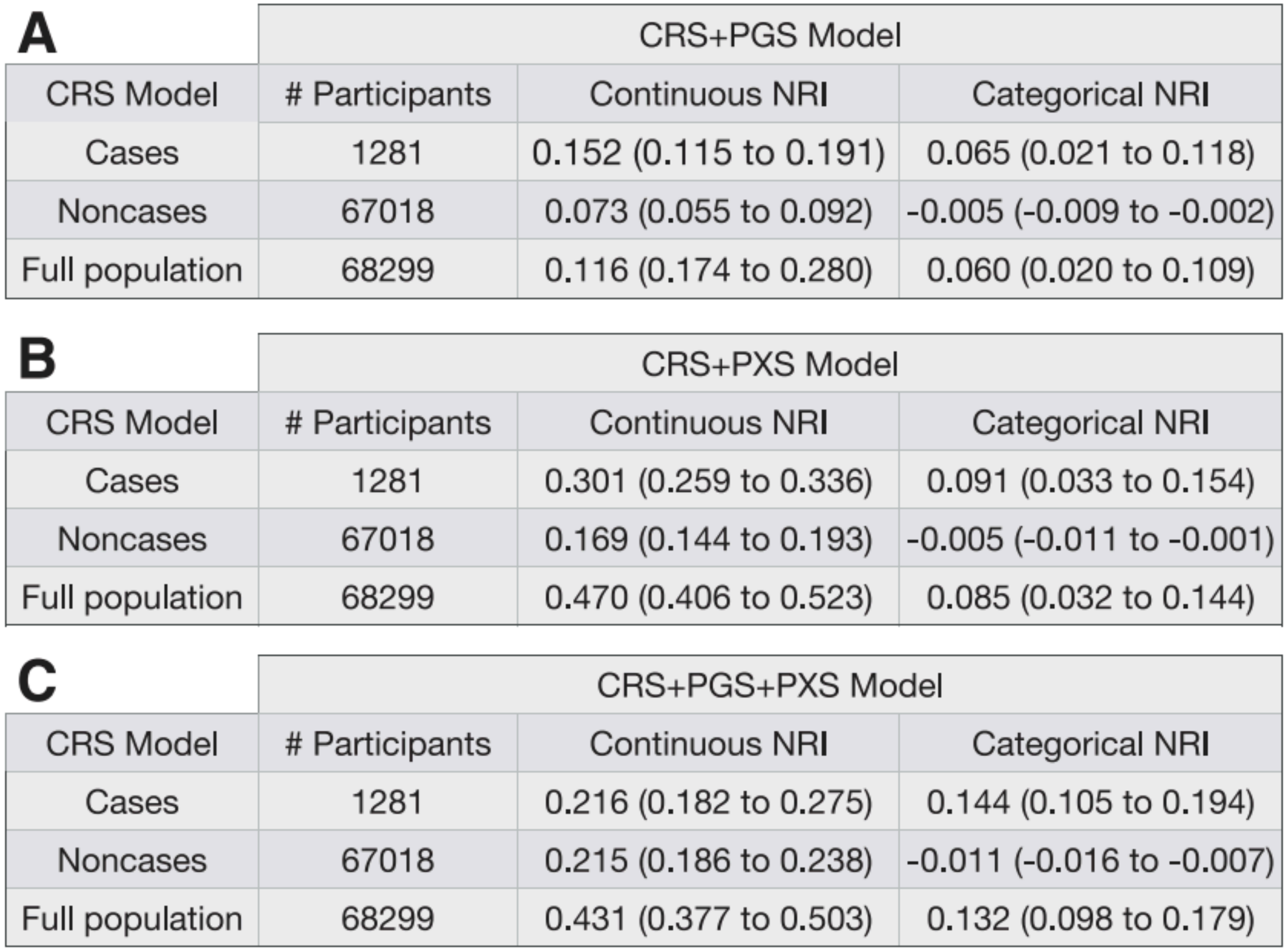
	C-Statistic (95% CI)		
	All	**Male**	**Female**
N	68299	32657	35642
# of Events	1281	844	437
Sex+Age	0.670 (0.656, 0.684)	0.629 (0.612, 0.646)	0.637 (0.612, 0.662)
PGS*	0.709 (0.696, 0.722)	0.680 (0.663, 0.697)	0.705 (0.682, 0.728)
PXS*	0.762 (0.749, 0.775)	0.732 (0.716, 0.748)	0.774 (0.753, 0.795)
CRS*	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



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)

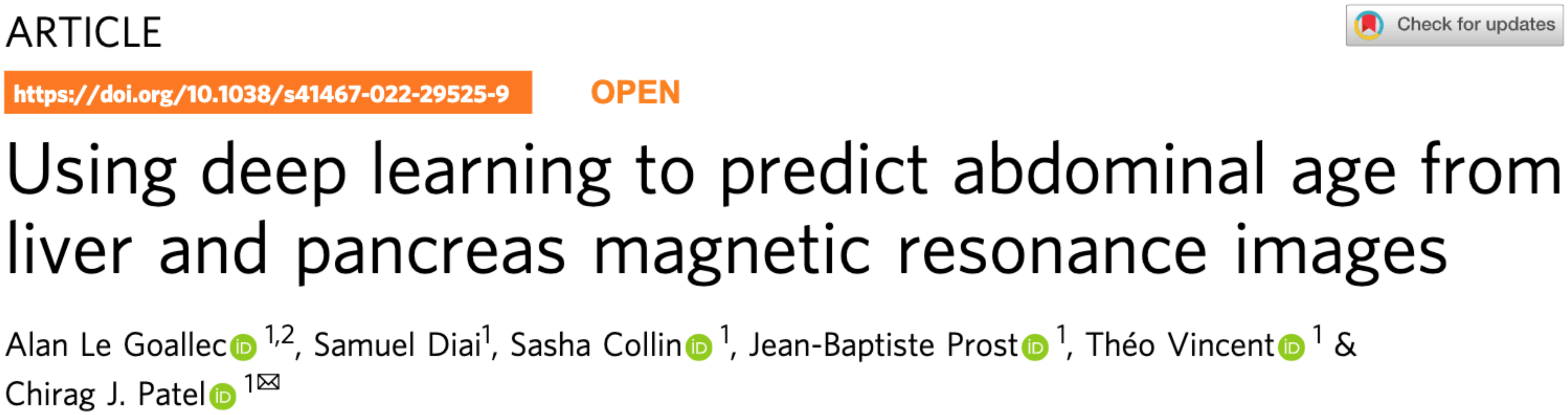
C	CRS+PGS+PXS Model		
CRS Model	# Participants	Continuous NRI	Categorical NRI
Cases	1281	0.216 (0.182 to 0.275)	0.144 (0.105 to 0.194)
Noncases	67018	0.215 (0.186 to 0.238)	-0.011 (-0.016 to -0.007)
Full population	68299	0.431 (0.377 to 0.503)	0.132 (0.098 to 0.179)

(see also Elliott et al, JAMA 2020)

Undiagnosed Diabetes
(A1C > 6.5%)
PRS: 0.696 (0.688, 0.705)
PXS: 0.756 (0.748, 0.764)

Diabetes Care 2021

Moving beyond glucose and BMI to dissect the multidimensionality of DM risk phenotypes: *Is it possible to predict pancreas and liver age?*



ARTICLE

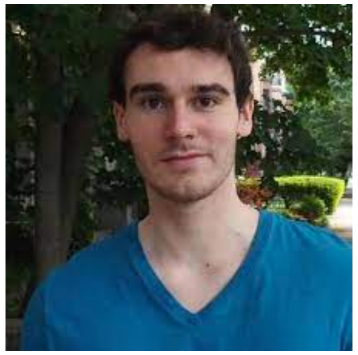
Check for updates

https://doi.org/10.1038/s41467-022-29525-9

OPEN

Using deep learning to predict abdominal age from liver and pancreas magnetic resonance images

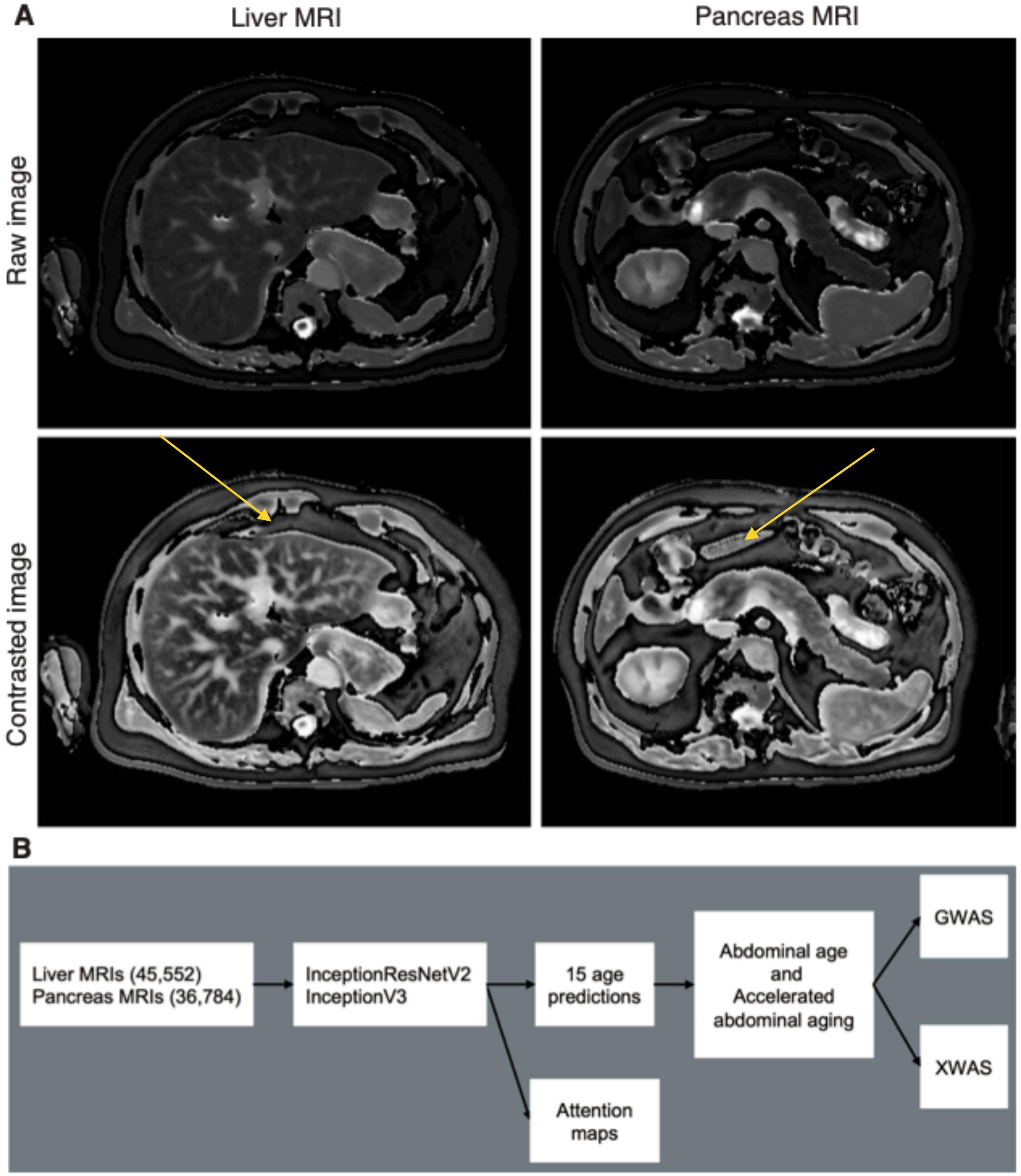
Alan Le Goallec[1,2], Samuel Diai[1], Sasha Collin[1], Jean-Baptiste Prost[1], Théo Vincent[1] & Chirag J. Patel[1]✉



Alan Le Goallec

Le Goallec et al, *Nature Communications* 2022

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



Liver MRIs (45,552)
Pancreas MRIs (36,784)

InceptionResNetV2
InceptionV3

15 age predictions

Attention maps

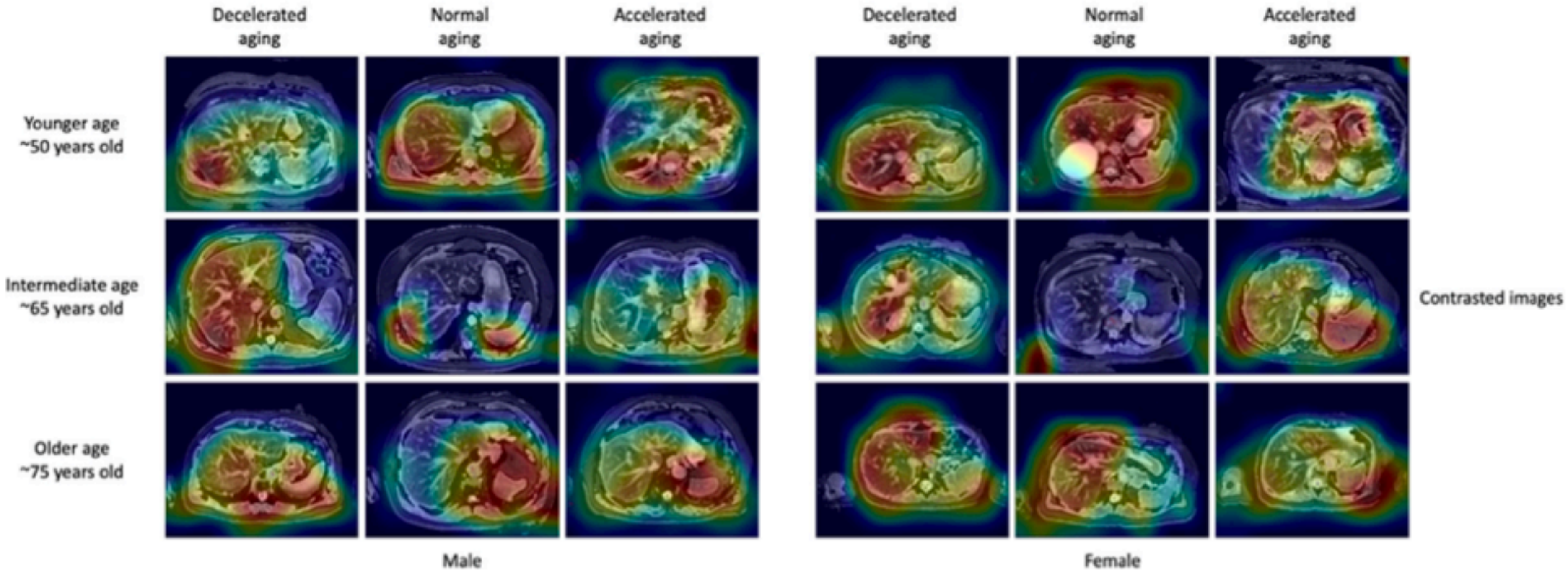
Abdominal age and Accelerated abdominal aging

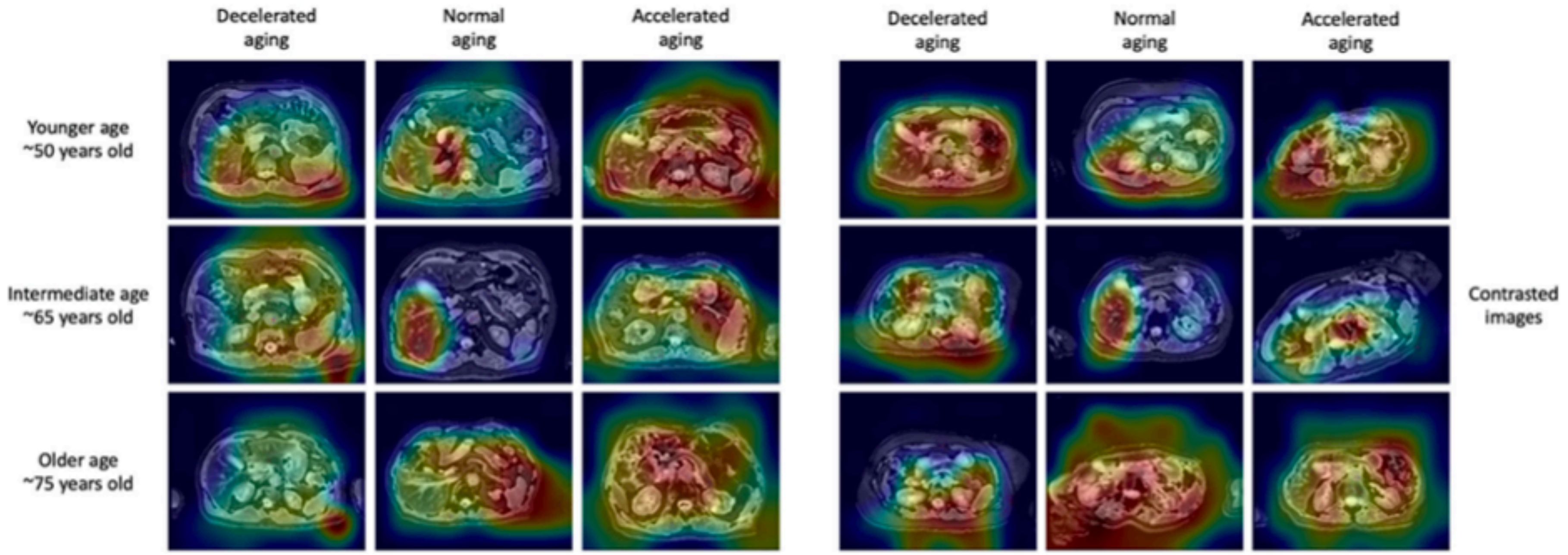
GWAS

XWAS

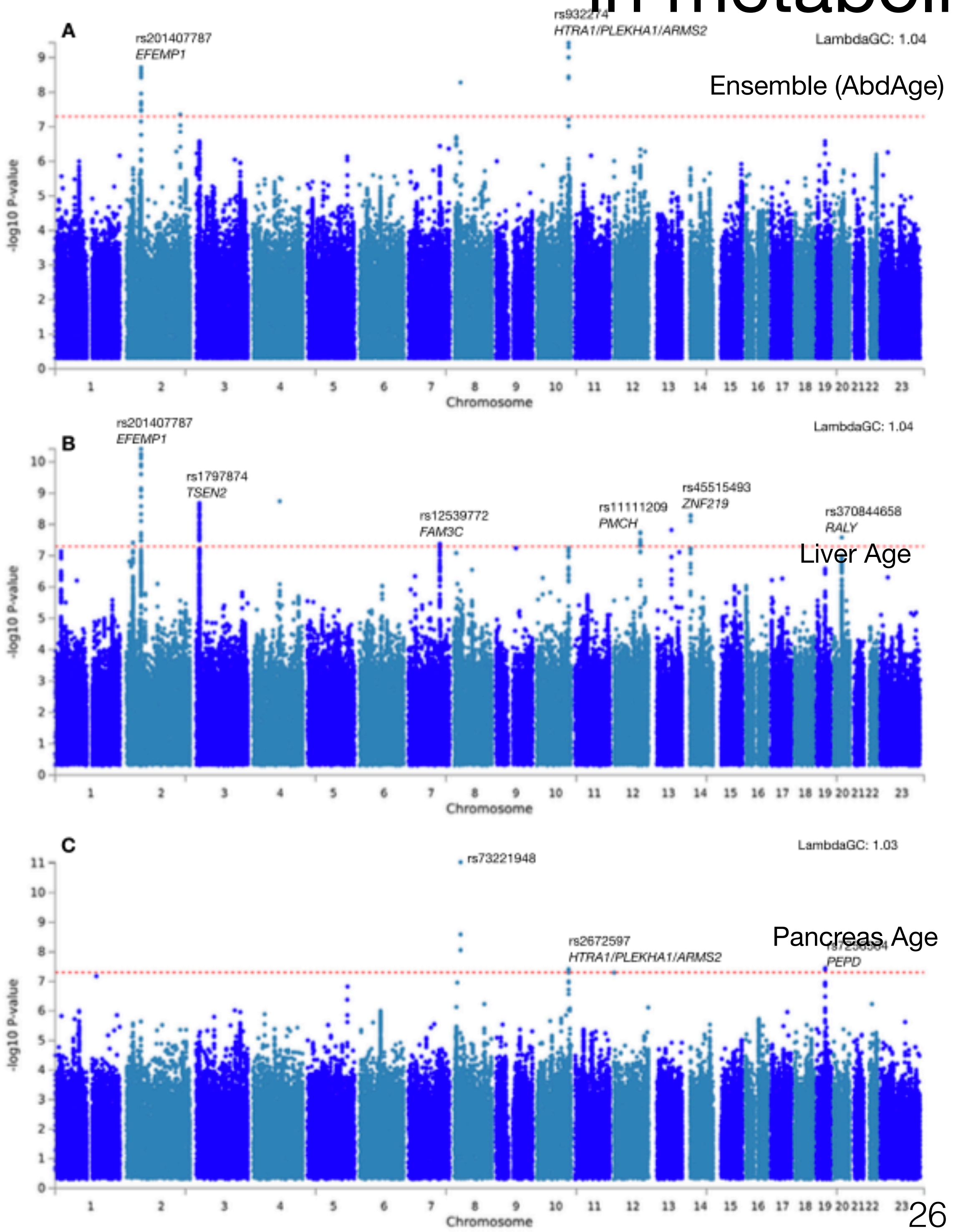
Le Goallec et al, *Nature Communications* 2022

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





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



Ensemble (AbdAge)

Liver Age

Pancreas Age

Genetic correlation between pancreas and liver: 0.86

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

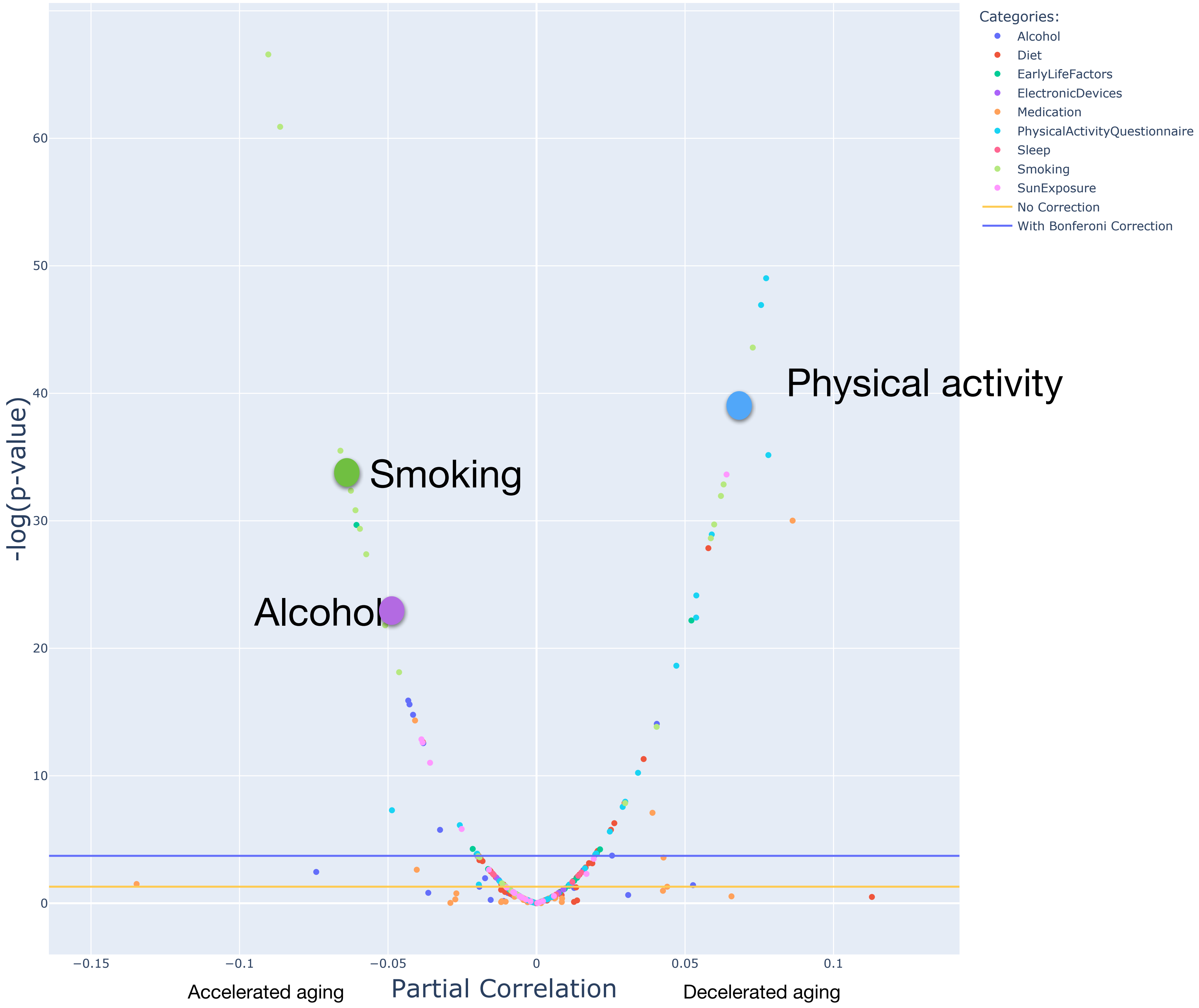
EFEMP1 (liver) is implicated in age-related macular degeneration

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

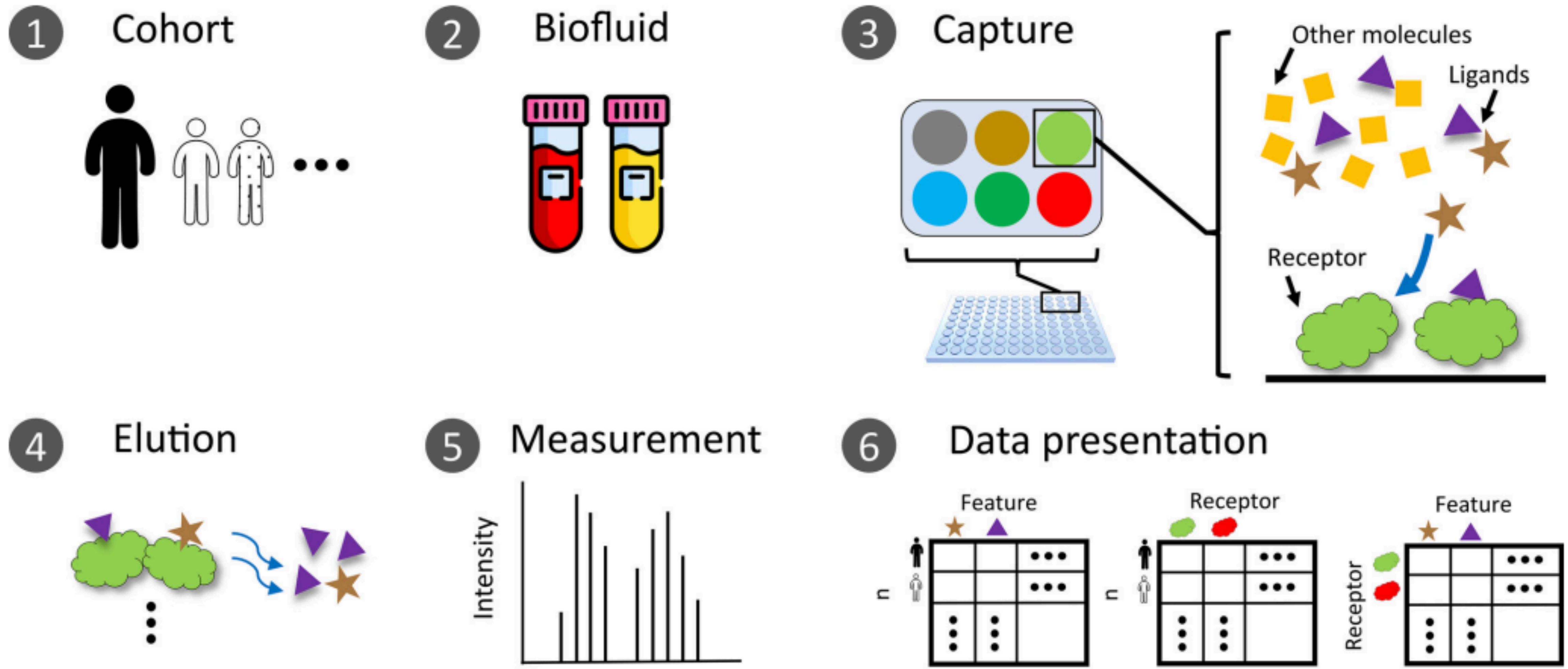
Le Goallec et al, *Nature Communications* 2022

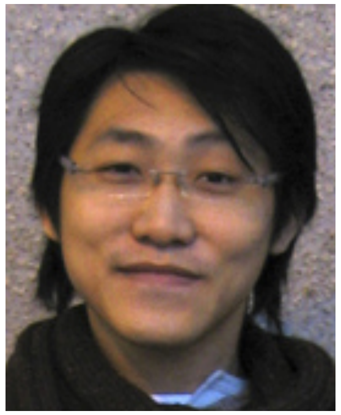
https://t2d.hugeamp.org/downloads.html

EWAS of Non-genetic/exposome factors (m=266) in abdominal aging: smoking, diet, physical activity, and alcohol



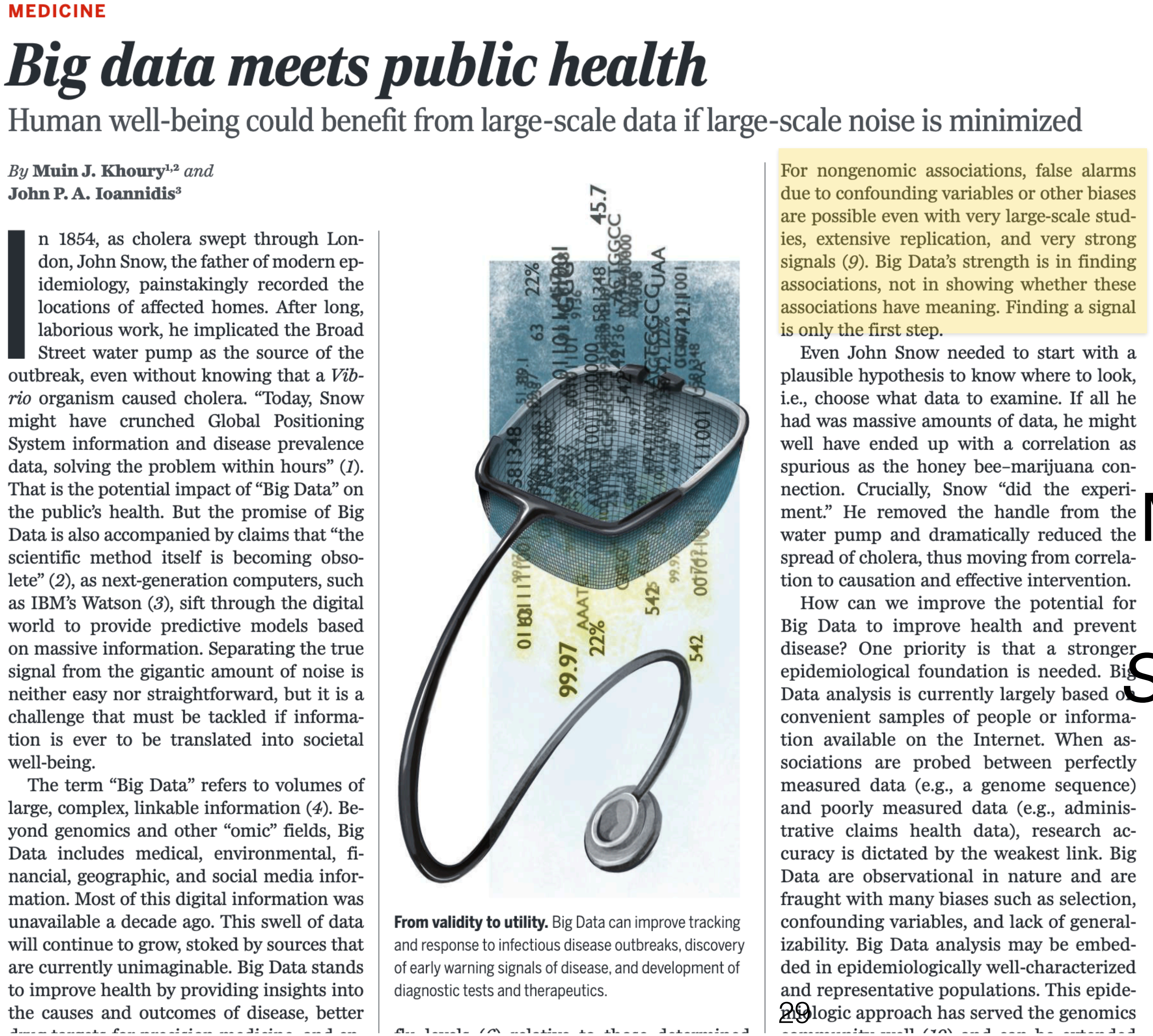
Biobanked samples to perform “functional” ***EWAS*** for discovery of biologically relevant exposures





Jake Chung

Chung et al., *Environmental Health Perspectives* 2021



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.

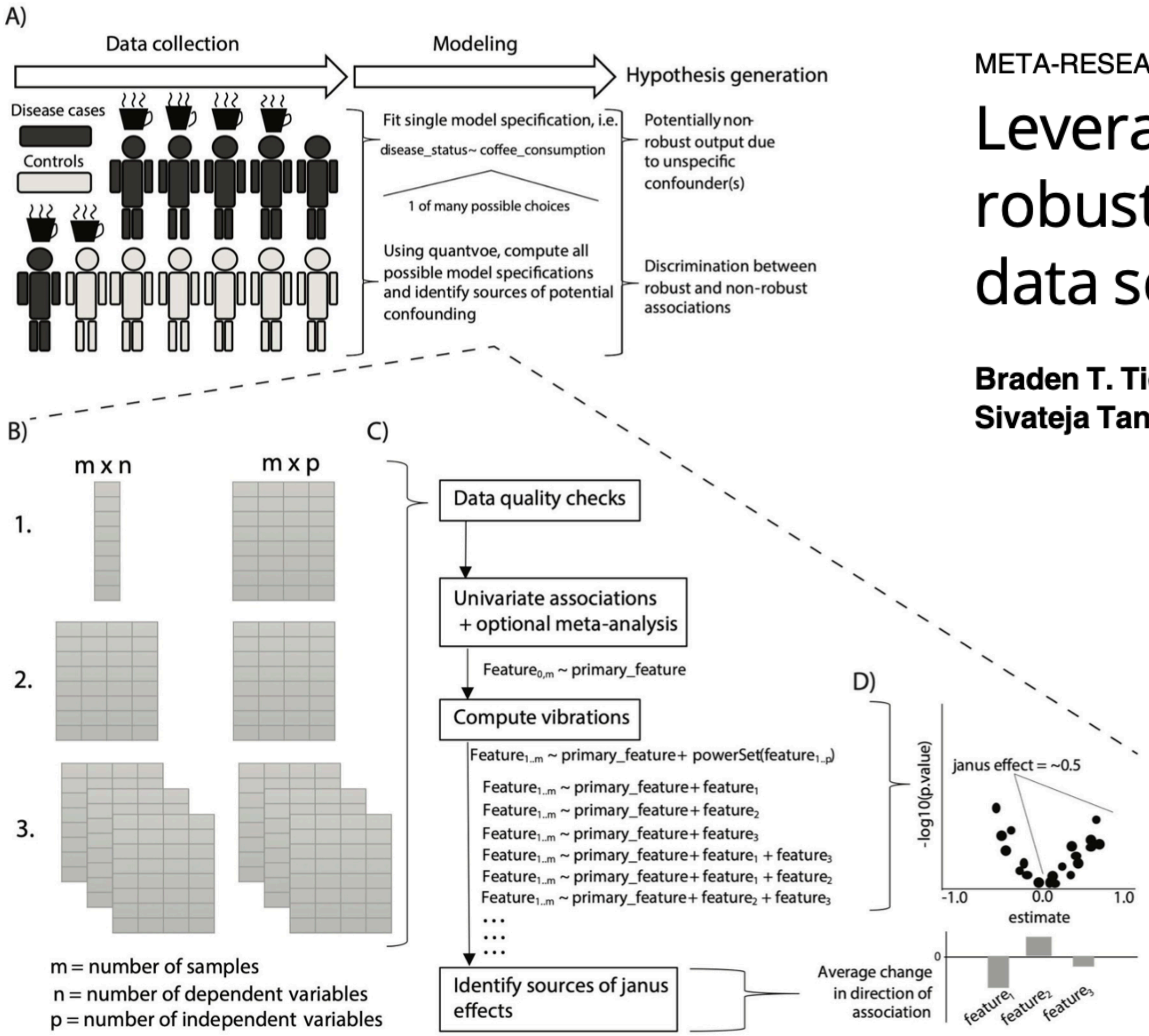
How can we improve the potential for Big Data to improve health and prevent disease? One priority is that a stronger epidemiological foundation is needed. Big Data analysis is currently largely based on convenient samples of people or information available on the Internet. When associations are probed 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

Confounding
Reverse causation
Missing information

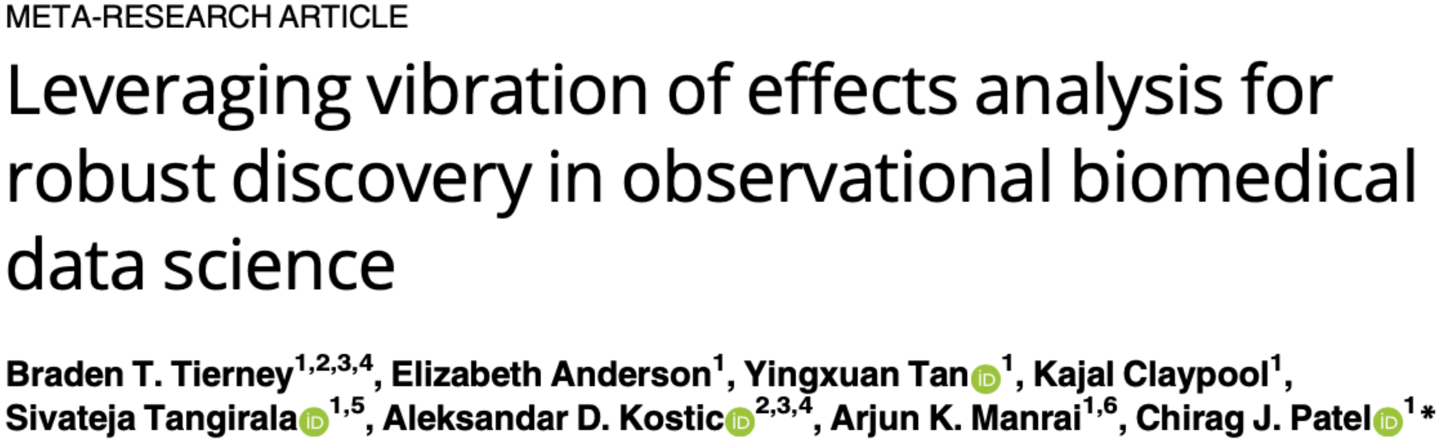
Sensitivity Analyses?
Triangulation?
Meta-analyses?

Science 2015

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



A)
Data collection
Modeling
Hypothesis generation
Disease cases
Controls
Fit single model specification, i.e.
disease_status~ coffee_consumption
1 of many possible choices
Using quantvoe, compute all possible model specifications and identify sources of potential confounding
Potentially non-robust output due to unspecific confounder(s)
Discrimination between robust and non-robust associations
B)
m x n
m x p
1.
2.
3.
m = number of samples
n = number of dependent variables
p = number of independent variables
C)
Data quality checks
Univariate associations + optional meta-analysis
Feature0,m ~ primary_feature
Compute vibrations
Feature1..m ~ primary_feature+ powerSet(feature1..p)
Feature1..m ~ primary_feature+ feature1
Feature1..m ~ primary_feature+ feature2
Feature1..m ~ primary_feature+ feature3
Feature1..m ~ primary_feature+ feature1 + feature3
Feature1..m ~ primary_feature+ feature1 + feature2
Feature1..m ~ primary_feature+ feature2 + feature3
...
Identify sources of janus effects
D)
janus effect = ~0.5
-log10(p.value)
-1.0
0.0
1.0
estimate
0
Average change in direction of association
feature1
feature2
feature3



META-RESEARCH ARTICLE

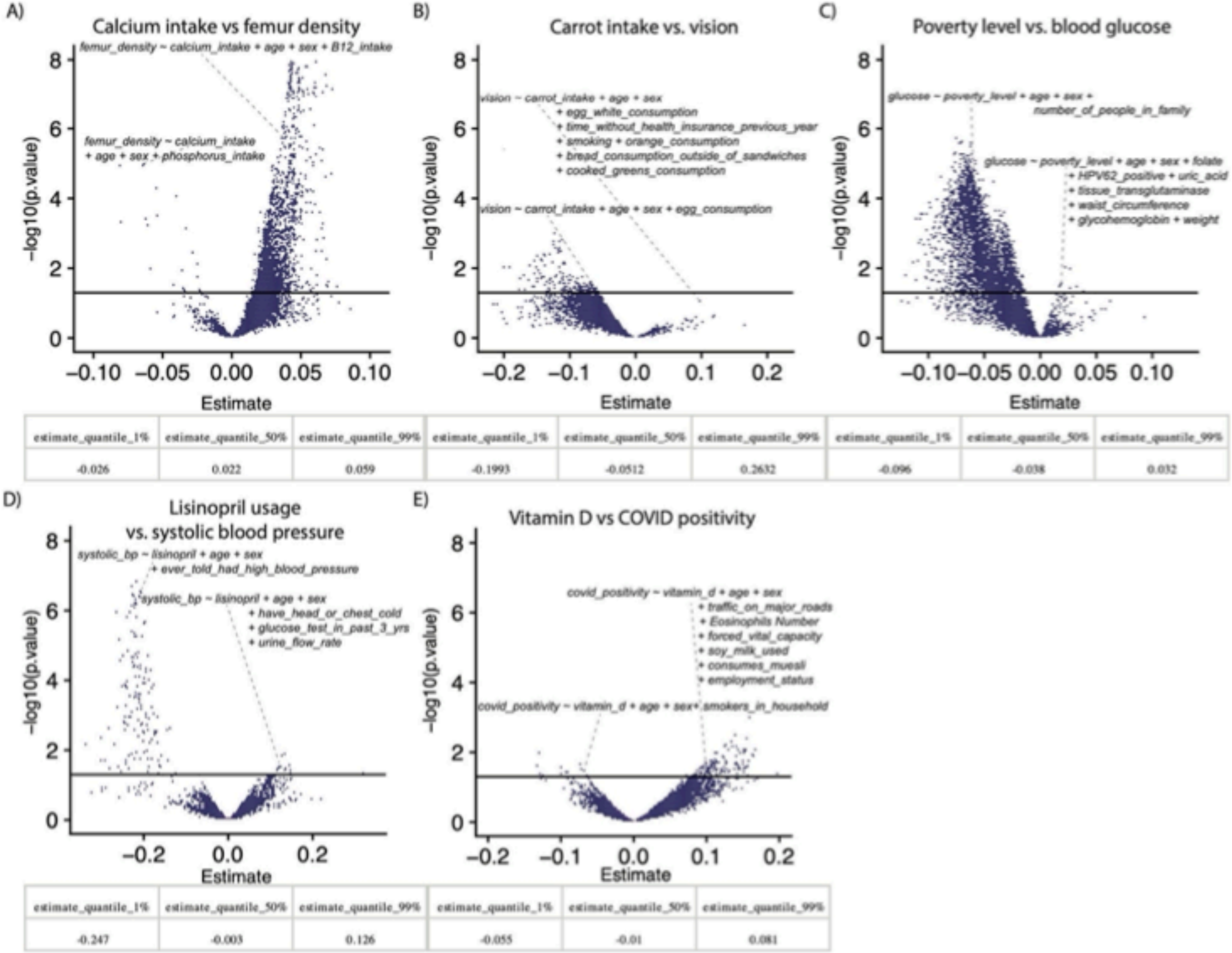
Leveraging vibration of effects analysis for robust discovery in observational biomedical data science

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



estimate_quantile_1%	estimate_quantile_50%	estimate_quantile_99%	estimate_quantile_1%	estimate_quantile_50%	estimate_quantile_99%	estimate_quantile_1%	estimate_quantile_50%	estimate_quantile_99%
-0.026	0.022	0.059	-0.1993	-0.0512	0.2632	-0.096	-0.038	0.032

estimate_quantile_1%	estimate_quantile_50%	estimate_quantile_99%	estimate_quantile_1%	estimate_quantile_50%	estimate_quantile_99%
-0.247	-0.003	0.126	-0.055	-0.01	0.081

Fig 2. Examples of VoE for prominent associations. Each point in the density plots represent at least one model. x-Axes are estimate size for beta-coefficients of interest (e.g., for panel A, the coefficient on the calcium intake variable). Quantiles show the range of estimate sizes for each above relationship. The y-axis is the −log10(p.value) of that association. The solid line is nominal ($p < 0.05$) significance. Data underlying these plots are available at https://figshare.com/account/home#/projects/120969 and S2 Table. VoE, vibration of effects.

It is possible to identify new and established exposures associated with health in big ***biobanked*** data!

(1) Discover & replicate new exposures and genes;
(2) Interrogate new biological pathways;
(3) “Triangulate” possible causal relationships;
(4) Perform meta-analysis and synthesis

All of Us℠ | The Precision Medicine Initiative®
THE FUTURE OF HEALTH BEGINS WITH YOU

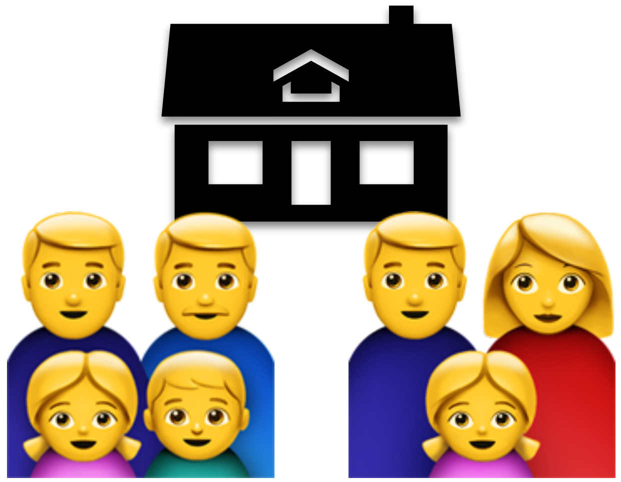
lifelines

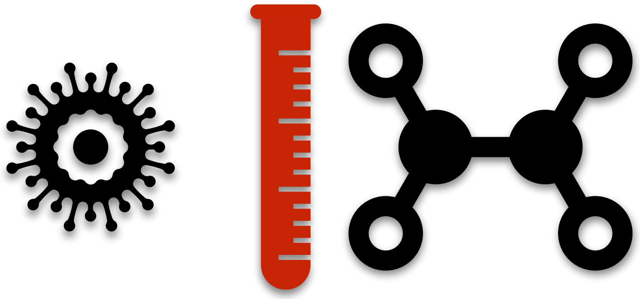
biobank uk

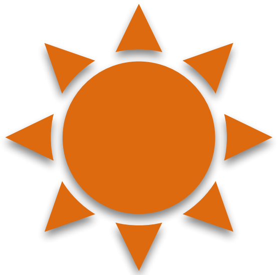
… what about to study early development and children?

Requirements for biobanked study of children to identify exposures associated with health and disease

- Consent (Kilma et al, *Genetics in Medicine* 2014)
- Measurement of development-relevant phenotypes
- Frequent measure through *in-utero* and developmental time
- Biosamples to assay the ***exposome***
- Linkages to health information of mom & dad
- Associations with future health outcomes (adolescence, adulthood)
- ***Geospatial*** exposome biomarkers: climate and air pollution
- Data approaches to harmonize across cohorts for meta-analyses and systematic reviews







Acknowledgements

RagGroup

Yixuan He
Jake Chung
Chirag Lakhani
Kajal Claypool
Tom Van Der Meer
Alan LeGoallec
Braden Tierney

Mentioned Collaborators

John Ioannidis
Peter Visscher
Arjun Manrai

NIH National Institute on Aging

NIEHS National Institute of Environmental Health Sciences

NSF

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