Discovering inter-individual differences in *exposomic* risk at biobank scale

Chirag J Patel

Catalyzing development and use of Novel Alternative Methods
NAMs Working Group, National Institutes of Health
8/21/2023

HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics

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

It is possible to identify new and established exposures and indicators of inter-individual variation at ***biobank scale***

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

lifelines

biobank uk

- Observational/non-randomized
- Convenience (self-selected)
- Single cohort and sample

- "Hypothesis-free"
- Biosamples
- Location of participants

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

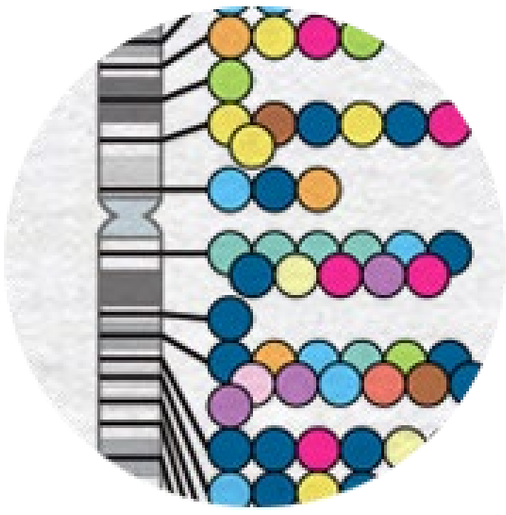
3,955 publications (as of 4/21/19)
136,287 *G-P* associations

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

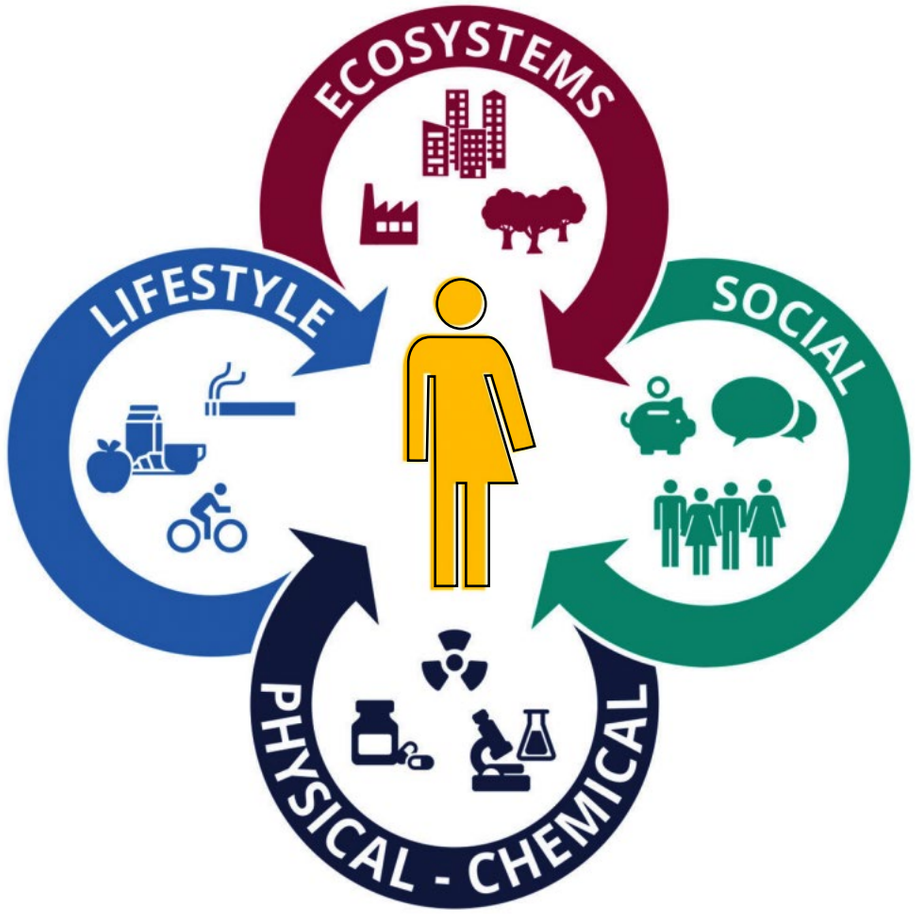
- Scaled for discovery
- Robust associations
- Negligible confounding
- Zero reverse causality
- *Little prediction capability*



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

Abdellaoui et al, *AJHG* 2023

The exposome: systematic exposures across domains & modalities



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

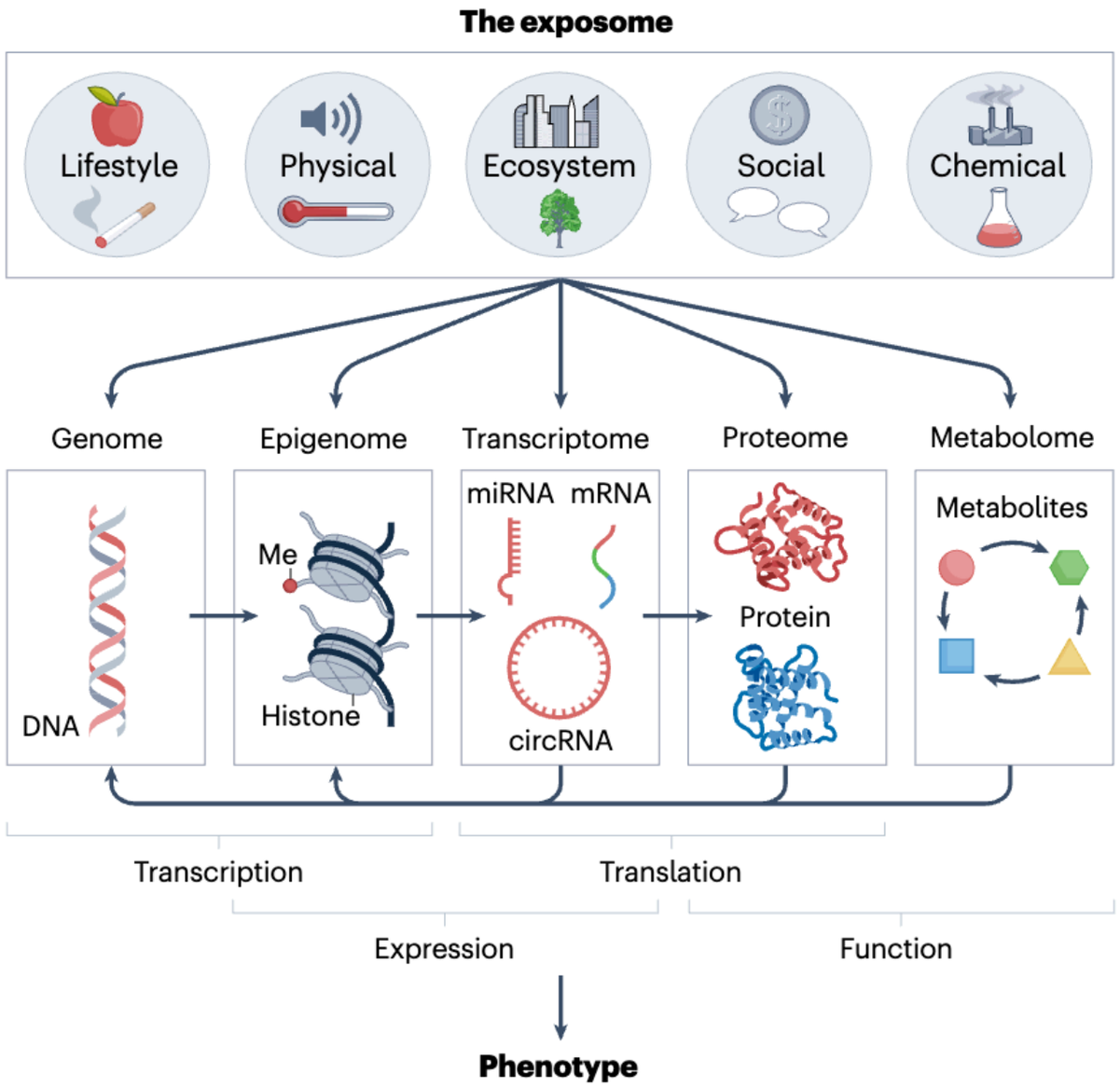
LIFESTYLE
- Physical activity
- Sleep behavior
- Diet
- Drug use
- Smoking
- Alcohol use

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

PHYSICAL - CHEMICAL
- Temperature/Humidity
- Electromagnetic Fields
- Ambient Light
- Odor & 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
- Illustration - Utrecht University

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

If the exposome is causal, they must induce some biological response: relating the exposome to biology and disease?!



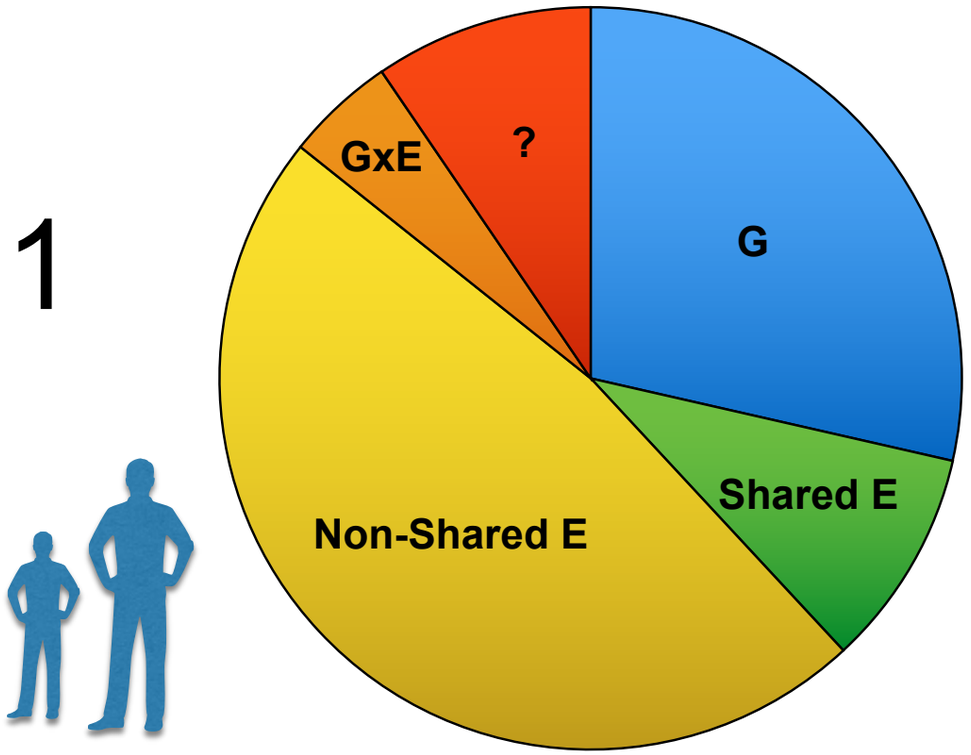
Are changes in 'omic response related to disease?

Wu et al, *Nature Reviews Genetics* 2023

Key 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):

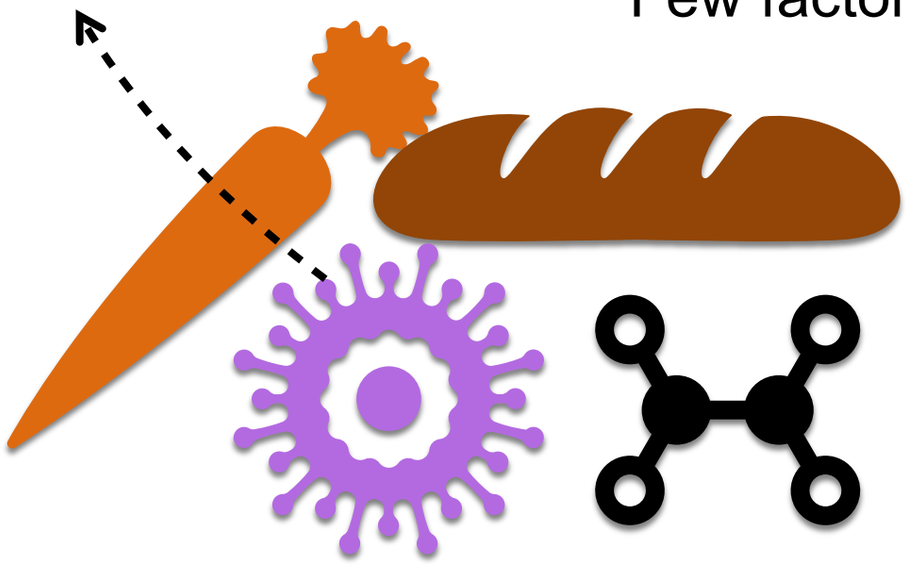
What factors are associated?

How do the exposures “add” up in aggregate?

Multiple factors of small effects?

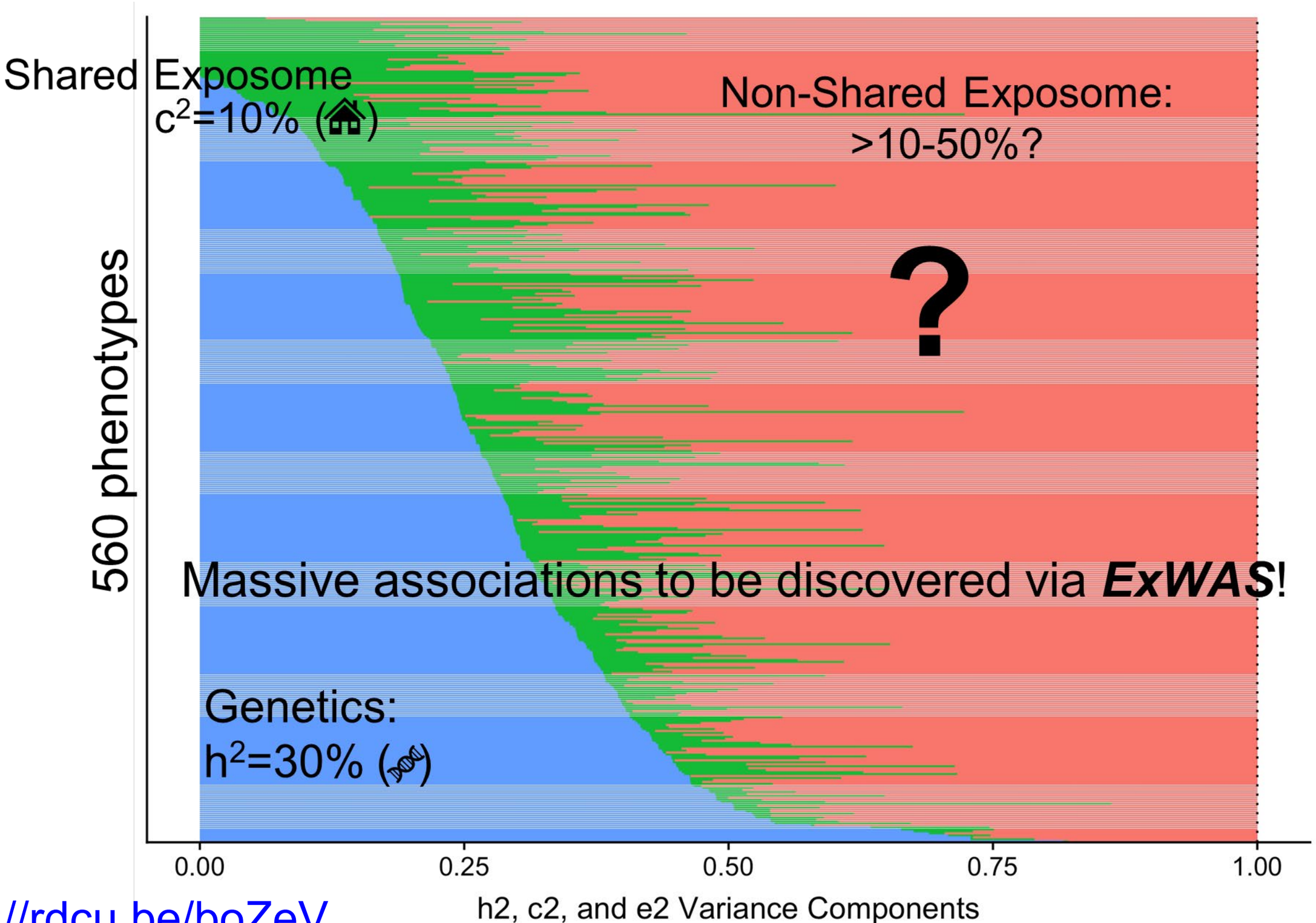
Few factors with large effects?

2



?

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



https://rdcu.be/boZeV

http://apps.chiragjpgroup.org/catch

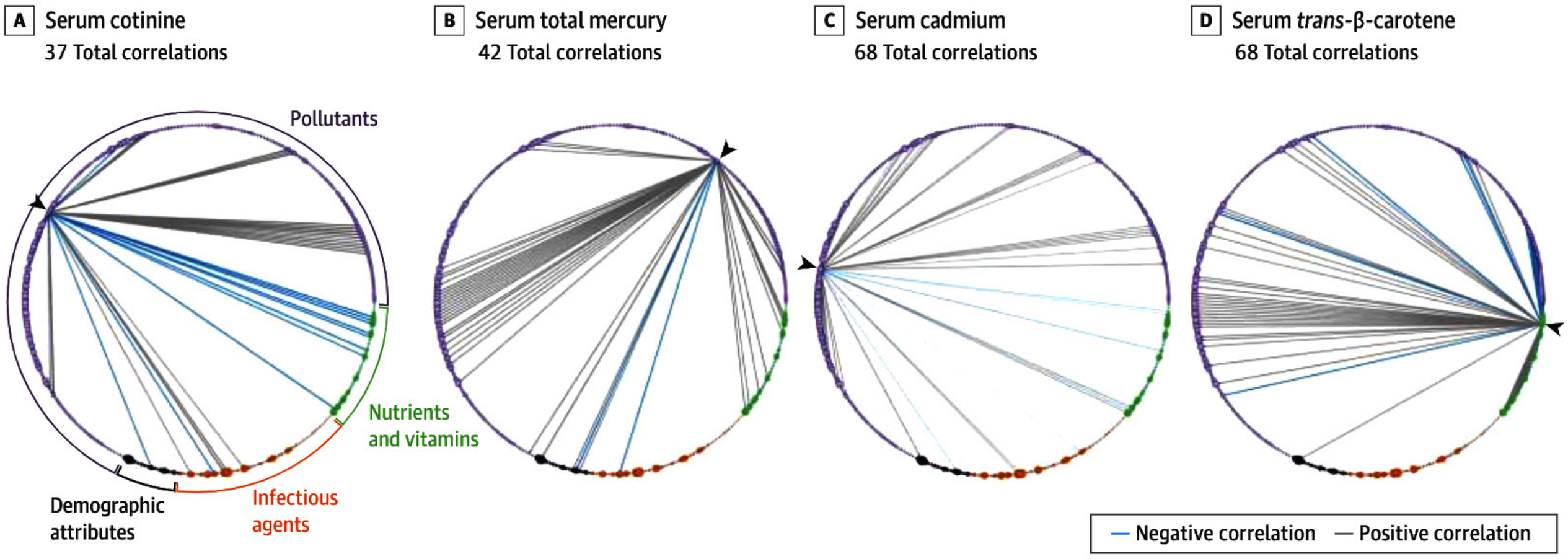
Lakhani et al., Nature Genetics 2019

How to measure the missing variation?
Modalities of the ***exposome*** in biobanks: diverse, time-dependent, and hierarchical

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

Modalities of the exposome: highly correlated - 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

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
R^2 (SBP): <1%

82 *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

IJE 2017

461 *E* and *P*

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]✉

Nature Communications 2020

374 *E*

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,*]

Env Int 2018

128 *E*

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

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

Diabetes Care 2021
UK Biobank

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

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]

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

Diabetes Care 2022
Personalized Environment and Genes (PEGS) cohort

A PXS may have utility those at highest aggregate risk or for reclassification of a “clinical risk score” - famhx, lipids, glucose, blood pressure

A

CRS+PGS 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)

B

CRS+PXS 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)

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

Bradford Hill is a mainstay for systematic assessment of exposures in disease for decision making

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)

At this first meeting of the Section and before, with however laudable intentions, we set about instructing our colleagues in other fields, it will be proper to consider a problem fundamental to our own. How in the first place do we detect these relationships between sickness, injury and conditions of work ? How do we determine what are physical, chemical and psychological hazards of occupation, and in particular those that are rare and not easily recognized?

I have no wish, nor the skill, to embark upon a philosophical discussion of the meaning of `causation'. The 'cause' of illness may be immediate and direct, it may be remote and indirect underlying the observed association. But with the aims of occupational, and almost synonymously preventive, medicine in mind the decisive question is whether the frequency of the un-desirable event B will be influenced by a change in the environmental feature A. *How* such a change exerts that influence may call for a great deal of research. However, before deducing `causation' and taking action we shall not invariably have to sit around awaiting the results of that research. The whole chain may have to be unraveled or a few links may suffice. It will depend upon circumstances.

Disregarding then any such problem in semantics we have this situation. Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation ?

Proc R Soc Med 1965

Bradford Hill criterion for exposome/genome are violated or do not apply: what role can NAMs have in re-articulating a heuristic?

1) **Strength of association (high risk)** ✗
   Most individual exposures might be small, some large

2) **Consistency of association** ✓
   Replicated in multiple cohorts?

3) **Specificity of association** ✗
   One exposure ~ one phenotype

4) **Temporality** ?
   Exposure comes before phenotype

5) **Biological gradient** ?
   Higher the exposure, the higher the risk or phenotype

6) **Biological plausibility** ?
   Does the mechanism have some prior?

7) **Coherence** ?
   Can the association be reproduced experimentally

Triangulation and meta-science: choosing different approaches that make different assumptions

Repeating experiments is not enough

Verifying results requires disparate lines of evidence — a technique called triangulation. **Marcus R. Munafò** and **George Davey Smith** explain.

*Is **smoking** causally linked to **outcome**?*

Cor(smoking, outcome)
Cor(smoking taxation, outcome prevalence)

Munafo and Davey Smith, Nature 2018

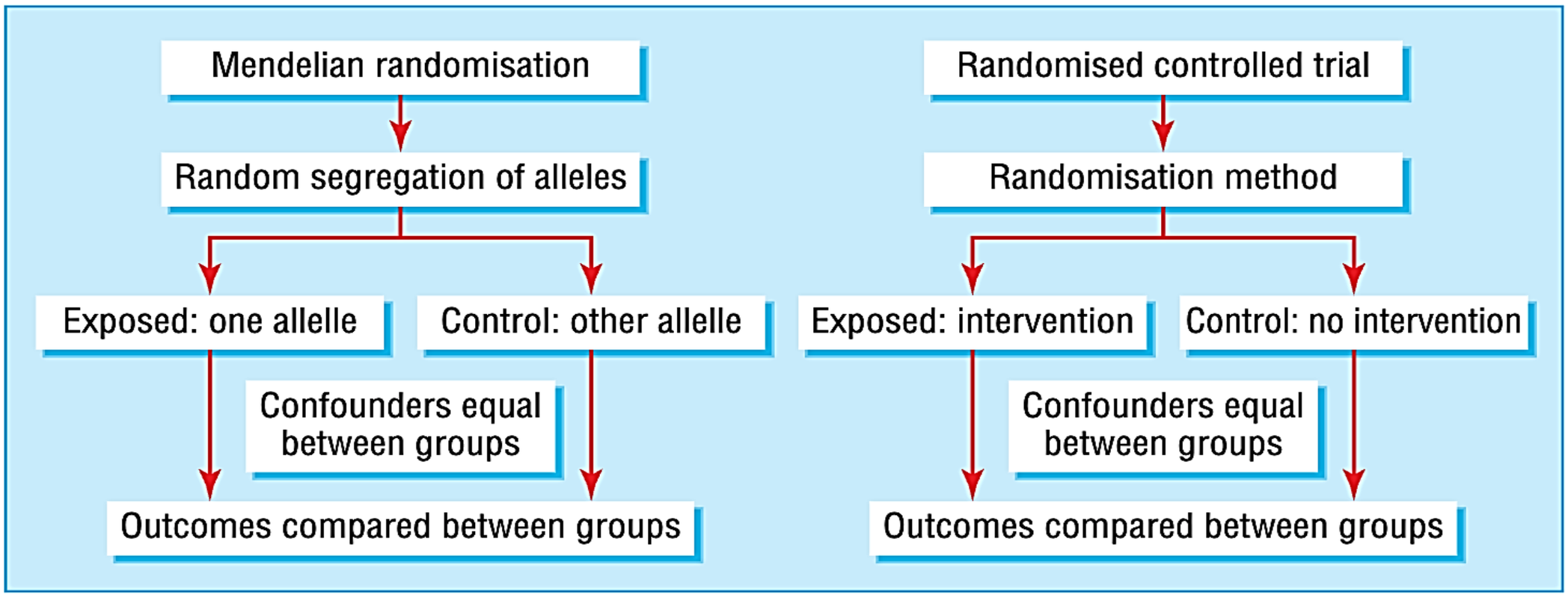
Mendelian randomization:
Genetic inter-individual variation as a “randomized trial”

What can mendelian randomisation tell us about modifiable behavioural and environmental exposures?

George Davey Smith, Shah Ebrahim

Using genetic variants as a proxy for modifiable environmental factors that are associated with disease can circumvent some of the problems of observational studies

BMJ 2005



Comparison of design of mendelian randomisation studies and randomised controlled trials

Guiding questions for capturing exposomic variability at biobank scale: from public health to biology

- **Real world surveillance:** what “new” and “old” exposures are prevalent in the population? How are they correlated?
- **High-throughput epidemiology at biobank scale:**
  - Are our current heuristics relevant for NAMs, genome, and exposome?
  - Do exposures influence phenotypic readouts (‘omics) in relevant tissues?
  - And are those multi-omic readouts associated with disease and clinically/biologically relevant?
- **Phenotyping:** How to map multiple and correlated exposures encountered in biobanks to disease-relevant NAMs

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

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

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