Published in final edited form as:
JAMA. 2014 June 4; 311(21): 2173–2174. doi:10.1001/jama.2014.4129.

Studying the Elusive Environment in Large Scale

Chirag J. Patel, PhD
Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center, Department of Health Research and Policy, Department of Medicine, Stanford University School of Medicine, Stanford, California, Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, California, and Meta-Research Innovation Center at Stanford (METRICS), Stanford, California

It is possible that more than 50% of complex disease risk is attributed to differences in an individual's environment.[1] Air pollution, smoking, and diet are documented environmental factors affecting health, yet these factors are but a fraction of the “exposome,” the totality of the exposure load occurring throughout a person's lifetime.[1] Investigating one or a handful of exposures at a time has led to a highly fragmented literature of epidemiologic associations. Much of that literature is not reproducible, and selective reporting may be a major reason for the lack of reproducibility. A new model is required to discover environmental exposures associated with disease while mitigating possibilities of selective reporting.

To remedy the lack of reproducibility and concerns of validity, multiple personal exposures can be assessed simultaneously in terms of their association with a condition or disease of interest; the strongest associations can then be tentatively validated in independent data sets (eg, as done in references 2 and 3).[2,3] The main advantages of this process include the ability to search the list of exposures and adjust for multiplicity systematically and report all the probed associations instead of only the most significant results. The term “environment-wide association studies” (EWAS) has been used to describe this approach (an analogy to genome-wide association studies). For example, Wang et al[4] screened more than 2000 chemicals in serum to discover endogenous exposures associated with risk for cardiovascular disease.

There are notable hurdles in analyzing “big” environmental data. These same problems affect epidemiology of1-risk-factor-at-a-time, but in EWAS their prevalence becomes more clearly manifest at large scale. When studying hundreds and thousands of exposures, tens and hundreds of associations often emerge that pass conventional statistical thresholds. Yet most of these seemingly statistically robust associations are correlates only, not causal

Corresponding Author: John P. A. Ioannidis, MD, DSc, Stanford University, 1265 Welch Rd, MSOB X306, Stanford, CA 94305 (jioannid@stanford.edu)..

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

associations. Reverse causality and confounding may underlie most of the observed strong correlations.

Based on the enormous number of potential interrelated correlations between multiple environmental exposures (depicted by edges in the Figure), it is uncertain whether there was ever any reasonable hope for traditional epidemiology to use rational thinking, biological plausibility, or some other reasoning to select and document risk exposures one at a time. For example, smoking (measured by cotinine levels) is clearly harmful, but it is also correlated with dozens of other exposures (Figure, A). Seemingly harmful associations of these exposures with diverse health outcomes may simply be attributable to their correlation with smoking. Pollutants such as mercury (Figure, B) or cadmium (Figure, C) may have multiple correlations with diverse seemingly “healthy” nutrients and other exposures. Moreover, any intervention that tries to influence one exposure node may inadvertently influence many others that are correlated. For example, from the EWAS vantage point, intervening on β-carotene (Figure, D) seems a futile exercise given its complex relationship with other nutrients and pollutants.

Given this complexity, how can studies of environmental risk move forward? First, EWAS analyses should be applied to multiple data sets, and consistency can be formally examined for all assessed correlations. Second, the temporal relationship between exposure and changes in health parameters may offer helpful hints about which of the signals are more than simple correlations. Third, standardized adjusted analyses, in which adjustments are performed systematically and in the same way across multiple data sets, may also help. This is in stark contrast with the current model, whereby most epidemiologic studies use single data sets with out replication as well as non–time-dependent assessments, and reported adjustments are markedly different across reports and data sets, even those performed by the same team (different approaches increase validity but must be reconciled and assimilated).

However, eventually for most environmental correlates, there may be unsurpassable difficulty establishing potential causal inferences based on observational data alone. Factors that seem protective may sometimes be tested in randomized trials. The complexity of the multiple correlations also highlights the challenge that intervening to modify 1 putative risk factor also may inadvertently affect multiple other correlated factors. Even when a seemingly simple intervention is tested in randomized trials (affecting a single risk factor among the many correlations), the intervention is not really simple. In essence what is tested are multiple perturbations of factors correlated with the one targeted for intervention. This means that randomized trials of interventions on putative protective environmental exposures (eg, diet or lifestyle) should be repeated in diverse populations for which the interrelated correlations might be different, before considered widely generalizable.

The EWAS model can be extended and improved. To capture time dependence, investigations must accommodate measurement of multiple environmental exposures at different times in the lifespan, particularly in development, and new analytical methods must be able to capture the complex temporal relationship between multiple exposures and future disease risk.[5] Second, little is known about how environment interacts with the genome. The current literature on gene-environment interaction is highly fragmented, nonsystematic, and

subject to selective reporting, suggesting the need for interdisciplinary gene-environment-wide association studies.[6] In addition, quality of measurements will dictate the breadth of any environmental research effort. However, quantitative and inexpensive methods to measure many environmental factors in a high-throughput manner (unlike genetic chips) are lacking. Mirroring the evolution of genomic measurements, this may change.

High-throughput ascertainment of endogenous indicators of environmental exposure that may reflect the exposome increasingly attractattention, and their performance needs to be carefully evaluated. These include chemical detection of indicators of exposure through metabolomics, proteomics, and biosensors.[7] Eventually, patterns of high-throughput biomarker measurements would need to be connected with external sources of exposure, such as behaviors, diet, and the built environment to translate findings to meaningful correlates and potentially modifiable environmental factors and exposures for individuals.

Much can be done today in lieu of having the perfect comprehensive exposome “chip.” First, most observational epidemiologic studies already measure more than a handful of risk factors. For example, consumption of multiple nutrient factors can be determined from dietary instruments, and multiple serum/urine biomarker levels are often ascertained. All such variables can be associated with phenotypes and traits of interest and reported simultaneously using EWAS. Second, epidemiologic investigations, especially those publicly funded, should be deposited in the public domain to encourage both standards development and integrative studies, like the Databases of Genotypes and Phenotypes. In the genomics field, funding agencies and scientific journals mandate that US federally funded gene expression experiment data be deposited in public repositories such as the Gene Expression Omnibus. The repository has been instrumental in development of technology for measurement of gene expression, data standardization, and reuse of data for discovery. Just as with the Gene Expression Omnibus, an “Exposure Omnibus” will help enable more powerful exposure-phenotype studies, assimilating data from around the world.

Further, there needs to be a common dictionary for environmental exposure. Such a dictionary would document how different exposures can be measured (eg, assay methodology), where they can be measured (eg, in urine, serum, self-reported), their source (eg, food, water, air, or consumable/industrial by-product), and prevalence (eg, who is exposed). A common dictionary would enable consistent classification of data and interoperability of different cohorts and promote data sharing. Information standards such as the PhenX toolkit,[8] a reference of standard ways to assess a few common environmental risk factors, are the beginning of such efforts. With an information infrastructure in place and tools such as EWAS, it is possible to build a search engine for environmental exposures while leveraging existing epidemiologic resources as new methods for measuring the environment emerge.

Acknowledgments

Dr Patel reported that he is supported by Career Award K99 ES023504 from the National Institute of Environmental Health Sciences and by a Fellowship Career Award from the Pharmaceutical Research and Manufacturers of America Foundation. Dr Ioannidis reported no disclosures.

REFERENCES

1. Rappaport SM, Smith MT. Epidemiology: environment and disease risks. Science. 2010; 330(6003): 460–461. [PubMed: 20966241]
2. Tzoulaki I, Patel CJ, Okamura T, et al. A nutrient-wide association study on blood pressure. Circulation. 2012; 126(21):2456–2464. [PubMed: 23093587]
3. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. PLoS One. 2010; 5(5):e10746. [PubMed: 20505766]
4. Wang Z, Klipfell E, Bennett BJ, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature. 2011; 472(7341):57–63. [PubMed: 21475195]
5. Buck Louis GM, Sundaram R. Exposome: time for transformative research. Stat Med. 2012; 31(22): 2569–2575. [PubMed: 22969025]
6. Thomas DC, Lewinger JP, Murcray CE, Gauderman WJ. Invited commentary: GE-Whiz! Ratcheting gene-environment studies up to the whole genome and the whole exposome. Am J Epidemiol. 2012; 175(3):203–209. [PubMed: 22199029]
7. Rappaport SM. Biomarkers intersect with the exposome. Biomarkers. 2012; 17(6):483–489. [PubMed: 22672124]
8. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX Toolkit: get the most from your measures. Am J Epidemiol. 2011; 174(3):253–260. [PubMed: 21749974]

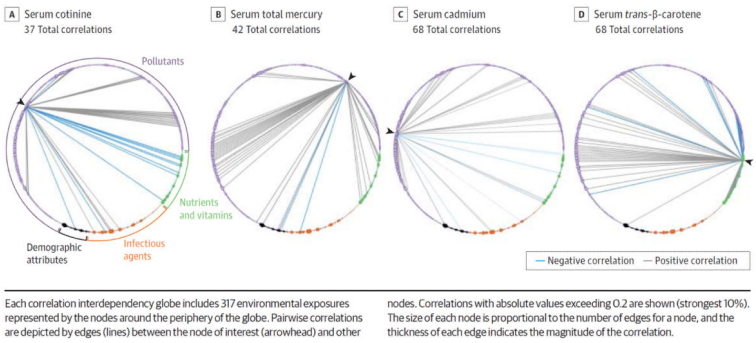

Figure.

Correlation Interdependency Globes for 4 Environmental Exposures (Cotinine, Mercury, Cadmium, *Trans*-β-Carotene) in National Health and Nutrition Examination Survey (NHANES) Participants, 2003–2004