Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2017 September ; 26(9): 1370–1380. doi: 10.1158/1055-9965.EPI-17-0459.

Opportunities and Challenges for Environmental Exposure Assessment in Population-Based Studies

Chirag J. Patel[1], **Jacqueline Kerr**[2], **Duncan C. Thomas**[3], **Bhramar Mukherjee**[4], **Beate Ritz**[5], **Nilanjan Chatterjee**[6], **Marta Jankowska**[2], **Juliette Madan**[7], **Margaret R. Karagas**[8], **Kimberly A. McAllister**[9], **Leah E. Mechanic**[10], **M. Daniele Fallin**[11], **Christine Ladd-Acosta**[11], **Ian A. Blair**[12], **Susan L. Teitelbaum**[13], and **Christopher I. Amos**[14]

[1]Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St, Boston, MA, 02115

[2]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA 92093

[3]Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90032

[4]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109

[5]Department of Epidemiology, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA, 90095

[6]Department of Biostatistics and Department of Oncology, Johns Hopkins University, Baltimore, MD 21205

[7]Division of Neonatology, Department of Pediatrics, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03755

[8]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, NH 03755

[9]Susceptibility and Population Health Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709

[10]Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, NIH, Bethesda, MD 20892

[11]Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD 21205

[12]Center of Excellence in Environmental Toxicology and Penn SRP Center, Perelman School of Medicine, University of Pennsylvania Philadelphia, PA 19104, United States; Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania Philadelphia, PA 19104

[13]Department of Preventive Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Corresponding Authors: Chirag Patel, 10 Shattuck St, Boston, MA. 02115, 617 432 1195, chirag_patel@hms.harvard.edu. Christopher Amos, Williamson Translational Building - 3, 1 Medical Center Drive, Lebanon NH 03756, Christopher.I.Amos@Dartmouth.edu.

[14]Department of Biomedical Data Science,, Geisel School of Medicine at Dartmouth College, Lebanon, NH 03755

Abstract

Background—A growing number and increasing diversity of factors are available for epidemiological studies. These measures provide new avenues for discovery and prevention yet they also raise many challenges for adoption in epidemiological investigations.

Methods—We evaluate 1) designs to investigate diseases that consider heterogeneous and multi-dimensional indicators of exposure and behavior, 2) the implementation of numerous methods to capture indicators of exposure, and 3) the analytical methods required for discovery and validation.

Results—Case-control studies have provided insights into genetic susceptibility but are insufficient for characterizing complex effects of environmental factors on disease development. Prospective designs are required but must balance extended data collection with follow-up of study participants. Two phase designs are described. We discuss innovations in assessments including the microbiome, mass spectrometry and metabolomics, behavioral assessment, dietary, physical activity and occupational exposure assessment, air pollution monitoring and global positioning and individual sensors. The availability of extensive correlated data raises new challenges in disentangling specific exposures that influence cancer risk from among extensive and often correlated exposures.

Conclusions—New exposure assessments offer many new opportunities for environmental assessment in cancer development.

Impact—We describe and evaluate the state of the art for evaluating high dimensional environmental studies.

Introduction

Both genetic and environmental factors contribute to the etiology of complex diseases. It has been recognized that there has been an inequality in GxE research, with less technological development and attention to environmental exposures (1, 2). Identifying environmental factors could result in potentially modifiable targets to decrease risk of disease and to enhance understanding of disease pathobiology.

Thousands of environmental exposure and risk-related behaviors are potential targets for epidemiological investigations and GxE research. In the era of “high-throughput exposure biology”, the concept of the *exposome* has emerged to describe comprehensive assessment of the totality of one’s “exposure” to environmental factors. Geographic Information Systems (GIS) and personal-level sensors are also creating new opportunities for epidemiological discovery. Furthermore, technologies to capture the external environment, such as ambient monitors, have established a role in environmental investigation. High-throughput measurement technologies have inspired the concept of precision medicine, an approach to capture individual genetic variation (3) and environmental exposures to tailor therapeutics and diagnoses for individual patients. These approaches will likely provide a

better understanding of chronic low-dose effects of exposures which will probably be a major contributor in understanding GxE effects. Environmental exposures broadly represent a broad range of physical, chemical, and biological agents but this article will primarily focus on factors that are potentially modifiable in human populations.

There are many challenges in implementing these new technologies for epidemiological population-based and clinical observational research. Issues that must be considered include: 1) the development of study designs to interrogate disease in the context of heterogeneous and multi-dimensional indicators of exposure and behavior, 2) implementation of numerous methods to capture indicators of exposure at various exposure levels, 3) analytical methods required for discovery and validation. In this commentary, we review the challenges and opportunities that these current and new techniques pose in epidemiological research.

Part 1. Design of Studies in the Context of High-Content Measurements

Study Designs

The successes of recent genome-wide association studies (GWAS) to discover and replicate variants associated with disease and phenotype (4) have made it tempting to extrapolate that similar agnostic approaches could lead to the discovery of many environmental and behavioral causes of diseases. Design of potential environment-wide association studies (EWAS) may provide novel insights into risk factors for complex diseases but raises new challenges due to complex measurement error, correlations between exposures, temporal variation, and biases that can plague observational studies. Large scale and untargeted environmental epidemiologic studies will critically depend on selection of study designs that can minimize false positive findings while maintaining robust power for the detection of underlying causal effects.

Prospective or cohort studies are ideal for conducting epidemiologic studies of environmental exposures since environmental exposure factors often change over time. Thus, prospectively collected, repeated measurements will be critical for assessing disease risk associated with the long-term average level of exposures as well as with their dynamic profiles. The optimal study design for balancing the number of participants and number of repeated measurements will depend on underlying hypotheses of interest, the intra-class correlation of the exposures (5), and the relative cost of recruiting individuals and measuring the exposures, and the types of phenotypic outcomes (e.g., quantitative trait or time-to-disease) under investigation. When stored biologic samples are to be used from an existing cohort study or surveillance program for assessment of new biomarkers, it is important to understand how content of the samples (e.g., chemical exposure biomarkers or RNA) may degrade over time with respect to the biological tissue being stored [e.g., (6, 7)]. For rare diseases, a strategy may be to combine data from multiple cohorts as was performed for the Cohort Consortium Vitamin D Pooling Project of Rarer Cancers (8).

Case-control studies, which are central to disease-specific GWAS, face intrinsic challenges for studying effects from environmental exposures due to well-known sources of bias such as reverse causality. This is particularly true for any biomarker-based approaches, for which the disease itself may alter biomarker levels or result in a change in behavior in individuals

(9). Case-control studies still have broad utility for studying GxE interactions due to the growing resources to model environmental exposures (i.e., model retrospective exposure) as well as robustness of multiplicative interaction parameters to effects of selection bias and non-differential misclassification (10, 11).

Hybrid designs can be used to combine advantages of cohort and case-control studies. At Phase-I, investigators may first establish a large cohort for the participants of which biological samples and data on certain risk factors, i.e. that are relatively inexpensive to ascertain, will be gathered. At phase-II, samples can be then selected from the Phase-I cohort in ascertaining biomarkers and more detailed exposures that may be expensive to conduct for the entire cohort. Two popular hybrid designs include the case-cohort (12) and nested case-control studies (13, 14). In the case-cohort design, the phase-II sample consists of all cases that arise during the follow-up of the cohort and a random sample of individuals from the cohort. In the nested case-control study, the phase-II sample consists of all cases and a set of matched controls for each case drawn from the subset of cohort members still at risk at the time the case occurred. Use of pre-diagnostic biological samples may avoid reverse causality bias despite the use of case-control sampling for subject selection at phase-II. For assaying samples in the laboratory, however, careful design is needed for batching cases and controls in a balanced fashion to avoid differential misclassification arising due to technical variability associated with various laboratory and instrument conditions. Hybrid designs are routinely used by many existing large cohort studies, such as the National Cancer Institute Prostate Lung Colorectal and Ovarian (PLCO) trial (15), for conducting biomarker based studies.

Samples at phase-II can be selected based on disease history of the subjects observed in the cohort as well as information on surrogates of exposure or other risk factors, a variant approach known as a two-phase design (16). Originally proposed for studying the relationship between a rare disease and a rare exposure (17)—a situation where neither the standard cohort nor case-control designs is efficient—the two-phase design can also be used to collect more information on exposures, confounders, or modifiers than would be feasible in the main study. A study may collect data on a few main exposures of interest at phase-I and on a larger set of other risk factors, including potential confounders or modifiers of the main exposures of interest, at phase-II (17). Enriching of phase-II sample by subjects who experience a rare outcome, i.e. the cases in the study, or/and individuals that have a rare exposure profile, can greatly enhance efficiency of identifying both main effects and interactions (18–21). The matched analog of the two-phase design is counter-matching (21, 22), in which cases and controls are sampled from the first stage in a manner that ensures that each matched set is discordant for the exposure surrogate, thereby improving power for main effects and interactions (23). A crucial component of the two-phase design is an analysis that 1) combines the information from both the main and sub-studies and 2) uses one of several methods for reducing bias that would be introduced by sampling jointly on exposure and disease (16). It is this combination of information from the two parts that distinguishes this design from simpler main study and validation (or pilot) sub-studies discussed above, where they are treated separately.

In addition to the issues raised above, cancer and other chronic diseases often involve an extended interval between exposure and disease and the effects of extended exposures are often cumulative. Thus, indices of cumulative exposure (e.g., pack-years for tobacco smoking) are widely used as the predictor in modeling exposure-response relationships. However, various other time-related variables such as age at exposure, time since exposure, attained age, or duration or level (acute high vs. chronic low) of exposure may modify the exposure-response relationship (24–27). Collecting extensive exposure information over time in large cohort studies would be a gold standard to strive towards but there are competing issues of cost and invasiveness of collecting extensive exposure data. On the other hand, the advent of new personal monitoring devices raises the potential to passively collect detailed data on participants over extensive periods with minimal cost (28). However, there are barriers to broad scale implementation that include the cost of providing sensors to participants and the management of extensive data from cohorts. The All of Us consortium (https://allofus.nih.gov/about/scientific-opportunities), funded by the Precision Medicine initiative is seeking to implement whole genome analysis with the application of new sensing and environmental measurement strategies for a cohort comprising 1 million participants.

Part 2. Types of Traditional and Emerging Exposure Measurement Modalities

Several different modalities exist to assess environmental exposures (Table 1) [for Review (29)] which can include external measures (30), biomonitoring (31), and measurements of biological effect (32). These measurements may be classified as either “individual-level” (e.g., serum levels of heavy metals measured on each participant or self-reported diet), or “ecological-level” which is based on spatiotemporal information on individuals, such as zip code at a certain point of time or with respect to an event.

These measurements are heterogeneous in type, the tissue or sample assayed, per sample cost, the number of variables assessed per assay, and potential sources of error. This heterogeneity poses an operational challenge in a large-scale epidemiological investigation, such as in data collection, data processing (e.g., assessing detected values, considering skewed distributions), data harmonization, data integration, and data analysis to provide biologically and clinically relevant signals (see next section). The approaches described below are all emerging, but are yet to be fully adopted in large scale epidemiology studies because of perceived or real needs for further validation, cost constraints, or other challenges. We describe some of the strengths and considerations for using these methods.

Some investigators have called for a single conceptual definition of heterogeneous measurements of exposure called the “exposome” (2, 31). The exposome considers multiple exposures humans encounter from conception to death (33) simultaneously. Christopher Wild has divided the exposome into three domains, including the ‘general external’, the ‘specific external’, and the ‘internal’ (34). The general external exposome includes indicators of socioeconomic status, financial status, and stress. The specific external includes factors such as radiation, infectious agents, pollutants, diet, lifestyle factors, medical

interventions. The internal exposome consists of internally measured exposure and phenotypic factors, such as indicators of metabolism, microflora, and inflammatory markers. If the concept is to be successful as a tool for discovery of exposures in disease, the heterogeneity of data measures seen in Table 1 must be addressed in appropriate study designs (see above) and in analyses (see below). A few exposome research efforts are now underway. For example, the Children's Health Exposure Analysis Resource (CHEAR) is a program funded by the National Institute of Environmental Health Sciences (NIEHS) to advance understanding about how environmental exposures impact children's health (35). CHEAR is designed to expand the range and access of environmental exposures assessed in NIH-funded children's health studies -- such as untargeted and targeted mass spectrometry based assays (Table 1). In Europe, the Human Early-Life Exposome (HELIX) project is bringing together six existing birth cohort studies comprising 32,000 mother-child pairs to study the impact that a broad array of exposures has upon development and disease (36). In this U.S. the Environmental influences on child health outcomes program (ECHO) is developing methodologies for identifying early determinants of child health and disease by characterizing the early exposome.

Microbiome

One aspect of immune dysfunction/disease has focused on the intestinal and lung microbiome, and the associated health signatures (37–49); the implementation of the Human Microbiome Project (50) has contributed significantly to more comprehensive and larger microbiome investigations in human populations (51). Recently, standardized techniques have expanded the study of the microbiome into large-scale studies (52) (Table 1). Establishing norms in large cohorts is an important first step to enable links of multiple exposures to changes in the microbiome and ultimately to long term health outcomes (53). Collection techniques, laboratory protocols, microbial DNA extraction kits, and even sequencing platforms often vary from study to study, creating challenges for data pooling. Not only will it be important to understand sources of variability in the collection, processing, and analyses of the microbiome (54–56), but continued evaluation of evolving sequencing platforms will be necessary as well. For example, there are two main methods for collection of microbiome information. In the first, 16S ribosomal RNA is targeted and sequenced. The 16S sequence fragments are classified using off-the-shelf bioinformatics tools, into operational taxonomic units (OTUs), and these OTUs are analyzed to understand the presence of different microbiome organisms in a given sample. In the second, the entire community of the microbiome is sequenced (called "metagenomic sequencing"), which not only provides information on what types of organisms are present, but also their "functional" capability through the sequencing of genes that are expressed in the sample (57).

There are a few but impactful examples of microbiome investigations in humans that demonstrate the association in human disease, such as colorectal cancer, integrate careful control over sample collection and processing with analysis of outcomes. In one such investigation, In one such investigation, Kostic and colleagues performed 16S ribosomal sequencing of microbiome organisms in 95 matched pairs of colon cancer tumors versus adjacent non-affected colon sites (58). Their data-driven investigation implicated species of the genus Fusobacterium, enriched in tumor versus non-tumor sites. While provocative and a

demonstration of creation of hypotheses in the association these investigators found between *Fusobacterium*-associated sequences and colorectal cancer are subject to concerns about reverse causality and the mechanism of tumor growth; for example, it is entirely possible that these specific bacteria accumulate in tumor sites and tissue because of the cancer itself. We expect that future investigations of unrelated or unpaired individuals will need to harness the study designs described above to strengthen claims of direction association.

Targeted and untargeted mass spectrometry

One approach to measuring biomarkers of exposure – either the actual exposure level or proxy (e.g., metabolites) -- includes mass spectrometry technologies (Table 1). Mass spectrometry can fall into two platform technologies, “targeted” and “untargeted”. “Targeted” mass spectrometry platforms detect chemicals that are known *a priori* in human tissue and urine and can be both indicators of the internal exposome or external exposome, such as lead, cadmium, and mercury. “Untargeted” platforms that allow for high content measurements, but may sacrifice exact identify of the chemical (output is limited to mass spectra) and may have lower sensitivity than a targeted assay (59). An advantage of untargeted platforms is that they are “agnostic”, enabling discovery of associatiosn with chemical entities that may have not been anticipated before the investigation. However, chemical analytic follow-up is often required to identify the chemical structure that emerges from an untargeted assay. One application of both targeted and untargeted mass spectrometry technology is for *metabolomics* (60), which applies an untarged approach to comprehensively examine the set of small molecule metabolites in human tissue, or indicators of the internal exposome and then follows up findings with a targeted substudy.

Work led by Hazen et al (61) has been an example of success in data-driven discovery of an endogenous indicator of a dietary factor (Trimethylamine N-oxide [TMAO]) linked to heart disease. First, Wang and colleagues began with an untargeted metabolomics approach to screen >2000 small chemical metabolites (measured with liquid chromatography mass spectrometry) in 50 cases that had incident myocardial infarction versus 50 matched controls without history of cardiovascular disease. After replication in another independent cohort, they found 3 correlated chemical analytes associated with cases versus controls, including TMAO. After examining the association of TMAO specifically in a larger cohort (N=1,876) with incident cardiovascular disease, they executed several rounds of mouse model experiments to begin to elucidate the causal association between TMAO and cardiovascular-related phenotypes. In the process, they found that TMAO “enhanced atherosclerosis” in mice, and that the mouse *microbiome* played a key role in producing TMAO from specific dietary factors. Since this impactful study, the investigators have gone on to demonstrate that suppressing specific flora through antibiotics influences TMAO production, and second, fasting levels of TMAO play a role in cardiovascular disease risk (62)

Sensor-based measures and physical activity assessment

Physical activity is a well-known risk factor for chronic disease. Many large epidemiological studies are including research grade accelerometer devices to assess physical activity (63) that can avoid misclassification bias in self reporting (64) (Table 1). Devices can be worn on the hip, wrist, or thigh, and can assess second by second behaviors (including sleep quality)

and postures (e.g. standing) which may be independently related to health (65). Collecting accelerometer data over multiple days allows researchers to assess patterns of behaviors (time of day and variability across days) in new ways so that more precise activity prescriptions (e.g. how much, when, and what behavior) can be given (66, 67). It is also possible to assess physical activity and related behaviors using Global Positioning System (GPS) derived coordinates on individuals, which are now omnipresent on mobile phones. However, this information must be linked with other sources of information (e.g., air pollutant monitors) in addition to individual-level information by merging on spatiotemporal coordinates, a straightforward but non-trivial information technology exercise. Researchers have been using GPS trackers alone or in combination with other monitors to assess exposure to pollution, outdoor time, and time spent in locations, such as food locations and parks (68–70). Research grade devices can cost considerably more than mobile phones, but have the capability of measuring over shorter intervals, which may give research an opportunity to capture location of an individual in almost real-time. However, it is an outstanding challenge in how to represent high-density information in an epidemiological analysis.

Occupational exposures

Occupational exposure investigations have provided key insights into etiological factors influencing chronic diseases such as cancer and heart disease. For one example, despite well-known risks for bladder cancer and other cancers among chimney sweeps and among agricultural workers, high risks for cancers and cardiovascular disease remain among these workers (71, 72). Studies of occupational cohorts have played a key role in epidemiological research because i) members of occupational cohorts may be subjected to quantifiable exposures, ii) exposures are often of long duration and consistent exposure allowing assessments to be reliably obtained. A challenge in occupational analysis is the requirement to collect detailed information from the cohorts and the complex coding required to assemble a detailed exposure history (Table 1). Traditionally, occupational exposures are assessed through detailed questionnaires. For example, in agricultural worker health, exposure level is determined by asking participants (a) use and frequency of use of a particular pesticide, (b), types of crops grown, (c) dietary intake and lifestyle factors, among other variables (73). While traditional occupational exposures in industrialized countries have been reduced, larger populations of samples and data are needed to estimate effects from traditional exposures at lower levels; however, some common ergonomic and psychosocial exposures are more difficult to measure. Aggregating such data often requires integration of occupational exposure information across multiple studies. Validating the exposures with external chemical analysis provides an objective approach for integrating data across studies (74). However, chemical validation may only be possible if biosamples are collected proximally to exposure. For example, measuring pesticide levels in farmers may not be possible when they are most busy and most exposed. Further, there is opportunity to collect this information digitally, via smart phone or computer to facilitate dynamic and remote collection of information.

Emerging tools for dietary assessment

New technologies now allow detailed short term dietary questionnaires, such as recalls or records, to be self-administered making possible their use in large scale prospective studies. However, investigator and respondent burden, as well as cost are still important considerations. One available and affordable technology for assessment of diet includes the Automated Self-Administered 24-Hour Dietary Assessment Tool (ASA24) which enables collection of self-administered 24-hour recalls and records on all mobile devices (75–77). Commercial food record apps are available but these generally provide data directly to the consumer and lack data on validation and quality control, and often lack extensive food and nutrient databases and data files of interest to research. Emerging technologies include image-based mobile phone apps in which participants take images of foods (often before and after consumption) and the goal of these technologies is to both identify and estimate portion size with minimal participant burden (78). To date, none are available nor validated for large-scale epidemiologic studies. In addition to the collection of self-report dietary assessment instruments, it is highly recommended that at least a sub-study be conducted in which recovery biomarkers are collected to allow for analyses that adjust for measurement error (79).

Part 3. Analytic and Data Integration Challenges

A dense correlational web of environmental variables poses challenges in multiplicity and power

It is apparent that epidemiological studies today and in the future have or will measure hundreds to thousands of these new and traditional environmental behavior-related and biologic variables (80). In Part 1, we discussed existing study designs that can be harnessed in investigating a handful of exposures in disease and in part 2, we discussed the emerging and existing tools that are used or can be used to measure environmental exposures. In this section, we discuss outstanding challenges and opportunities to marry these existing study designs and new high-throughput measures to discover new exposures in disease.

The number of variables in today’s genome-wide investigations, which now can query tens of millions of variants in association with a phenotype simultaneously, have led investigators to explicitly address issues such as type 1 and type 2 error through rigorous multiplicity control and harmonizing across numerous populations to ensure power for discovery. However, as documented elsewhere, the burden of type 1 error and type 2 error increases in GxE investigations (81). Furthermore, when assessing multiple exposures simultaneously, a dense correlational web between exposures may make discerning true interactions with an exposure versus those induced by correlations with the other correlated exposures (GxE confounding) difficult.

Therefore, to address this explicitly, it is important to have an assessment of the prevalence and variation of multiple exposures of interest. Cross-sectional but representative studies, such as the National Health and Nutrition Examination Survey (NHANES) (82), which collects information on many health related factors, can be useful for characterizing the variability and co-variability of factors for multiple environmental exposure biomarkers (80,

83, 84). If there are sets of highly correlated exposures, then disentangling their individual effects will require studying them together in studies of very large sample size. On the other hand, data from highly correlated exposures can be combined using data reduction techniques to reduce the number of variables to be measured in a large scale epidemiologic study where the initial goal may be detection of association of disease with broad classes of exposures. We emphasize that when using these techniques, biological interpretation is fraught with difficulty and data reduction of multiple correlated exposures is but just a first step to understanding associations between a class of exposures and a phenotype. Further still, studies such as NHANES are useful, but must be expanded to include all facets of the population. For example, NHANES does not take urine from children under six years of age. Repeated cross-sectional measures of biomarkers of exposure may also provide a comprehensive view of the intra-class correlation, or measurement error, of existing and new assays.

Data-driven searches of environmental confounders associated with phenotypes are also possible with the variables presented in Table 1 and study designs discussed in the previous section. In fact, some investigators have executed “exposome/environment-wide studies” to search for and replicate exposure-phenotype correlations. We anticipate the same challenges exist for exposome-wide studies as for GWAS, such as multiplicity correction and power. However, harmonizing across “exposome” measurements for added power and replication (Table 1) remains a problem. Further still, exposure and behavior variables are densely correlated (80, 83, 85). For example, we (Patel) estimated pairwise correlations between 317 environmental exposures of participants of NHANES (80, 86). For example, serum cotinine (a metabolite of nicotine), total mercury, cadmium, and trans-b-carotene were correlated with 37, 42, 68, and 68 other exposure biomarkers. Given this number of potential correlates with these biomarkers of exposures, it remains a challenge to identify exposures that are causally related to a phenotype or other exposures (e.g., confounded) and assess mediation (e.g., one exposure coming before or after another). Reverse causation (e.g. the phenotype coming before exposures) can be addressed through longitudinal studies and repeated measures can provide insights into temporal trends. However, an outstanding challenge remains in how to interpret associations given a dense correlational web of multiple factors. Previously, we argued that an association between an exposure and phenotype needs to be interpreted differently depending on what other correlations exist (80, 87). Some more robust statistical methods that can filter for associations and jointly model effects from many correlated factors show promise to assist in model selection, but should be evaluated when the factors are heterogeneous in measurement.

Data management challenges and emerging cloud-based solutions

Managing large epidemiological cohort databases with both genetic and environmental information is not a straightforward task. First, recruiting and collecting biological samples and information from participants adequate for determination of environmental exposure that are compatible with the study design is a challenge. Extensive data are collected frequently, requiring large amounts of disk drive space and computer processors for computation (and often perhaps spread in multiple and differently formatted data files). The problem is amplified when trying to analyze data that is collected at high-frequency, such as

daily or hourly. Third, sharing of data and tools across investigator sites can also be a hindrance to data use.

Multiple solutions exist to address these challenges [for a review, see (88)] and genome-wide investigations provide examples that demonstrate these solutions. For example, standardization of data units, such as genetic variants, have enabled compatibility across studies and harmonization to increase power in genome-wide studies. Common data files to represent data, such as “variant call files” for genotypes have enhanced creation of analytic tools. Standardization of ways investigators measure and collect non-genetic data, through efforts such as PhenX (89), is one way forward to enable data compatibility.

Addressing computational-related challenges is becoming easier with advances in computer infrastructure, such as “elastic” cloud computing that provide on-demand access to computer resources (such as disk space, memory, and processing time for computer intensive calculations). These infrastructures are emerging as both commercial and academic-based solutions in this space. As of this writing, the National Institutes of Health have established a “Cloud Commons” program to enhance the procurement of cloud computer resources and software for NIH grantees (see: https://datascience.nih.gov/commons). The program specifically promises to provide tools and serves to access (1) cloud computer environments, (2) publicly available datasets, and (3) software services to enable investigators to provision computer resources and share data resources with others.

Integrating genetic factors with emerging environmental, behavior, and biologic variables in epidemiological investigations

Larger-scale gene-by-environment interaction analyses that consider millions and thousands of genetic and environmental variables are fraught with challenges. Further still, GxE analyses require care to manage the diverse measurement profiles of genetic data versus environmental exposure data. As we have written earlier (90), a purely data-driven search for interactions between G number of genetic variants and E number of environmental variables would require G x E possible tests. For example, given G = 1 million genetic variants (commonly measured on a GWAS array) and E= 100 environmental exposures results in up to 1 million times 100 individual hypothesis tests for interaction (100 million!). The multiple comparison burden for querying the large sample space is prohibitive and the sample size requirements (81) to achieve adequate power will number in the 10s of thousands if not much more. As touched on above, there are a number of ways to “trim” the search space to a priori selection of candidate genetic variants or environmental exposure factors, including (a) querying those that have strong main effects from GWAS or EWAS (91) and emerging analytic methods such as two-step approaches (see review in Gauderman), (b) use of alternate methods estimate of the false discovery rate (FDR) of putative signals (92), and (c) use of biological priors as described in (93) to select genotypes that have documented influence on changes in gene expression. One such database includes the “Genotype-tissue expression” (GTEx), which provides genetic variants that are linked to tissue-specific (e.g., blood, lung, brain) gene expression levels (94). We outline several heuristics in Patel (90).

Heterogeneity of study and measurement error

One of the principal challenges in large-scale exposure association studies, in contrast to recent GWAS where precise genotype measurements are usually available with advanced genotyping assays, is the ubiquitous presence of exposure measurement error (95, 96). In conducting large-scale, multi-center/cohort exposure association and extending to gene-by-environment (GxE) analysis, there are significant challenges with harmonization of exposure data across multiple cohorts and understanding differing levels of exposure heterogeneity across studies (1, 97, 98). This is further compounded by the possible existence of differences in exposure measurement error in different studies or a very commonly encountered situation when limits of detection for exposure biomarkers across studies can be different due to differences in the exposure assay technologies used by the investigators (Table 1). While most, if not all, epidemiological cohorts measure many variables on their participants (80), the current literature is mostly limited to reporting associations between a single or a handful of exposures with a handful of phenotypes within a single study. Development of new methods with multiple exposures in the consortium-based setting will be required to assess exposures and phenotypes that span different studies and populations all simultaneously to limit reporting biases and false positive reporting (99–101) and demonstrate an EWAS-type analyses [e.g. (102)].

As new instruments and technologies become available to measure exposures in novel ways, it is critical to conduct studies to understand the sources of variability in the underlying measurements. To assess between and within subjects' sources of variations, these studies should include both a sample of individuals from an exposed population and a sample of measurements within each individual. Measurements within a subject may include various types of replicates to assess technical variability of the instrument and temporal variation in exposures. A recent study (103), for example, examined sources of variation in measurement of a panel of 539 urinary metabolites using liquid chromatography-mass spectrometry (LC-MS) and gas chromatography-mass spectroscopy (GC-MS) using data generated from 17 male subjects with 2–3 samples per person spread over 2–10 days. High reliability (i.e. low within person variability) was observed for most the metabolites.

While there is an extensive literature on misclassification and measurement error in the statistical and epidemiological literature, almost all the published studies focus on effects on marginal associations; fewer papers study its effects exclusively on interactions. Some of the earlier literature in GxE studies in this area considers measurement error in both genes and exposures (104–108). The findings from these studies indicate that in general, under both differential and non-differential misclassifications in E, the estimate of the multiplicative interaction parameter will be biased towards the null. An important research direction, specific to the GxE context, has been to study the role of GxE association and exposure misclassification simultaneously (109–112). In the presence of external validation data with true gold standard exposure measures that allow for estimation of the exposure misclassification probabilities, methods for correcting for measurement error have been shown to lead to enhanced power (110, 113, 114). Internal reliability designs, where exposures or genotypes are measured twice on a subset of subjects (109), or exposure

enriched designs (115), can also be employed to correct for measurement error and increase power.

Multiplicity of possible interaction tests and replication challenges

New and larger numbers of environmental, microbiotic, and behavioral variables provide new "dimensions" in the space of possible GxE interaction tests (90). For example, current Genome Wide Interaction Studies (GEWIS) (95, 116) execute only one interaction test per genetic locus. With new exposure measures, the space of possible tests increases to G x E possible tests, where G is the number of genotypes (often >1M for common SNPs) and E is the number of exposure or behavior-related tests [see also(117)].

Power and multiple testing burden pose an almost insurmountable challenge in multiple hypothesis correction and power required to detect GxE. The recent review by Gauderman *et al* provides further details (81). Methods to execute GEWIS will need to be extended to prioritize pairwise GxE tests, such as through biological plausibility (118, 119) and/or analytic approaches, such as focusing on genotypes and exposure variables that have strong main effects (91, 120) or are prevalent in the population (e.g., present in over 10%). However, one issue that will remain includes assessing GxE in the face of measurement error described above. An alternative to model interactions includes using "genetic risk scores" (GRS) (121, 122) and, analogously, "environmental risk scores" (ERS) (123). These approaches collapse additive environmental and genetic main effects into a single variable. Then, the ERS and GRS are tested in interaction. While this mitigates the issue of multiple testing and is useful to estimate disease risk, identifying causative loci or exposure agents is not possible with this method. One compelling approach is to single out environmental or genetic factors that have strong a priori evidence from GWAS, EWAS, and/or prospective studies [e.g.(102, 124, 125)].

Finally, replication of findings, or assessment of concordant associations across independent samples, will require harmonizable measures between studies and sample sizes suitable to detect effects (to avoid "winner's curse" type associations) (126, 127). Often, identification of cohorts will be difficult given the heterogeneity of measurements (Table 1) and scarcity of resources. Creation of "database of databases" that document cohort resources or provide summary statistics across GxE tests will be one way to enable investigators to replicate findings.

Discussion

There are a growing number of measurement modalities that are now or soon will be accessible for use in epidemiological investigation. The promise of incorporating these measures includes discovering novel factors that may be useful for clinical prognostics, for prevention, or even explaining disease etiology. One example of the successful identification of a novel gene-environment interaction through high dimensional analyses is presented by the finding that relatively common variants of *CHRNA5* influence smoking behavior (128, 129) and lung cancer risk (130, 131). Further, more detailed analyses, of the impact of these variants on attributes of smoking behavior and tobacco cessation programs, showed that the genetic factor specifically affects time to smoking cessation and the finding that carriers of

at-risk variants benefit substantially from pharmacological intervention in smoking cessation, whereas non-carriers do not benefit (132–134). Most interactions cause a marginal effect on risk that can be identified from either a genome-wide association study or from an environmental assessment of risk. However, understanding the full impact of the joint effects of genetic and environmental exposures over time requires reconstruction of exposures and behaviors in the context of the specific genetic background of individuals. Identifying novel gene-environment interactions that were not detected initially by their marginal effects from either environmental or genetic exposures usually requires large sample sizes, which can be achieved in some cases by coordinated studies from existing cohort studies. The All of Us cohort (https://allofus.nih.gov/), recently funded as a part of the Precision Medicine Initiative, seeks to collect extensive environmental and multi-omic measures from 1 million participants over extended periods of time, towards understanding the interplay of genetic and environmental exposures over time. This large cohort study should allow novel gene-environment interactions to be identified.

Incorporation of measurement profiles may enable epidemiologists to explain “missing heritability” in common variant-phenotype associations through assessment of gene-by-environment/microbiome/behavior interactions. But ultimately, shaping public health policies for prevention may be the most important elements to yield from these new measures. One hope is that these new and current measures will enhance efforts in “precision medicine” by enabling better prediction of therapies as a function of both genetic and environmental factors. This future also opens opportunities to tackle new methodologic challenges.

Acknowledgments

Financial Support.

Support for this review has been provided by the following NIH grants. Funding from R00ES023504 and R21ES025052 supported Dr. C.J. Patel. Funding from R01CA17997 supported J. Kerr and M. Jankowska. Funding from P01CA1956569 supported D.C. Thomas. Funding from R01ES023541 and R21ES025573 supported B. Ritz. Funding from P30CA023108 supported M.R. Karagas, J. Madden and C.I. Amos. Funding from P20GM104416 and P01ES022832 supported M.R. Karagas and J. Madden. Funding from U01DD00046, R01ES025216 and R01ES025531 supported. M.D. Fallin. Funding from P30ES013508 and P42ES023720 supported I. Blair. Support from U2CES026555 supported S. Teitelbaum. Support from U01CA196386, R21CA191651 and R01CA186566 and GM103534 supported C. I. Amos.

References

1. Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM, Tank NCIG-ET. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. Genetic epidemiology. 2013; 37:643–57. [PubMed: 24123198]
2. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev. 2005; 14:1847–50. [PubMed: 16103423]
3. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015; 372:793–5. [PubMed: 25635347]
4. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012; 90:7–24. [PubMed: 22243964]

5. Barrera-Gomez J, Spiegelman D, Basagana X. Optimal combination of number of participants and number of repeated measurements in longitudinal studies with time-varying exposure. Stat Med. 2013; 32:4748–62. [PubMed: 23740818]
6. Hebels DG, Georgiadis P, Keun HC, Athersuch TJ, Vineis P, Vermeulen R, et al. Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. Environ Health Perspect. 2013; 121:480–7. [PubMed: 23384616]
7. Brimo F, Aprikian A, Latour M, Tetu B, Doueik A, Scarlata E, et al. Strategies for biochemical and pathologic quality assurance in a large multi-institutional biorepository; The experience of the PROCURE Quebec Prostate Cancer Biobank. Biopreservation and biobanking. 2013; 11:285–90. [PubMed: 24194978]
8. Abnet CC, Chen Y, Chow WH, Gao YT, Helzlsouer KJ, Le Marchand L, et al. Circulating 25-hydroxyvitamin D and risk of esophageal and gastric cancer: Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. Am J Epidemiol. 2010; 172:94–106. [PubMed: 20562192]
9. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. Am J Epidemiol. 1992; 135:1019–28. [PubMed: 1595688]
10. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet. 2001; 358:1356–60. [PubMed: 11684236]
11. Wacholder S, Chatterjee N, Hartge P. Joint effect of genes and environment distorted by selection biases: implications for hospital-based case-control studies. Cancer Epidemiol Biomarkers Prev. 2002; 11:885–9. [PubMed: 12223433]
12. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986; 73:1–11.
13. Thomas DC. Use of computer simulation to explore analytical issues in nested case-control studies of cancer involving extended exposures: methods and preliminary findings. J Chronic Dis. 1987; 40(Suppl 2):201s–8s. [PubMed: 3667867]
14. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. Am J Epidemiol. 1990; 131:169–76. [PubMed: 2403467]
15. Zhu CS, Pinsky PF, Kramer BS, Prorok PC, Purdue MP, Berg CD, et al. The prostate, lung, colorectal, and ovarian cancer screening trial and its associated research resource. Journal of the National Cancer Institute. 2013; 105:1684–93. [PubMed: 24115361]
16. Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. J R Stat Soc Series B Stat Methodol. 1997; 59:447–61.
17. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol. 1982; 115:119–28. [PubMed: 7055123]
18. Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. Lifetime Data Anal. 2000; 6:39–58. [PubMed: 10763560]
19. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. J R Stat Soc Series C Appl Stat. 1999; 48:457–68.
20. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. Am J Epidemiol. 1988; 128:1198–206. [PubMed: 3195561]
21. Langholz B, Borgan Ø. Counter-matching: A stratified nested case-control sampling method. Biometrika. 1995; 82:69–79.
22. Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. 1996:35–53.
23. Andrieu N, Goldstein AM, Thomas DC, Langholz B. Counter-matching in studies of gene-environment interaction: efficiency and feasibility. Am J Epidemiol. 2001; 153:265–74. [PubMed: 11157414]
24. Thomas DC. Pitfalls in the analysis of exposure-time-response relationships. J Chronic Dis. 1987; 40(Suppl 2):71s–8s.
25. Thomas DC. Models for exposure-time-response relationships with applications to cancer epidemiology. Annu Rev Public Health. 1988; 9:451–82. [PubMed: 3288239]
26. Hauptmann M, Pohlabeln H, Lubin JH, Jockel KH, Ahrens W, Bruske-Hohlfeld I, et al. The exposure-time-response relationship between occupational asbestos exposure and lung cancer in two German case-control studies. Am J Ind Med. 2002; 41:89–97. [PubMed: 11813213]

27. Crump KS, Allen BC, Howe RB, Crockett PW. Time-related factors in quantitative risk assessment. J Chronic Dis. 1987; 40(Suppl 2):101s–11s. [PubMed: 3667856]
28. Betts KS. Characterizing exposomes: tools for measuring personal environmental exposures. Environ Health Perspect. 2012; 120:A158–63. [PubMed: 22469821]
29. Grossblatt, N., editor. Committee on Human and Environmental Exposure Science in the 21st Century; Board on Environmental Studies and Toxicology; Division on Earth and Life Studies; National Research Council. Scientific and Technologic Advances. Exposure Science in the 21st Century: A Vision and a Strategy. Washington, D.C: The National Academies Press; 2012. p. 106-96.
30. Turner, MC., Nieuwenhuijsen, M., Anderson, K., Balshaw, DM., Cui, Y., Dunton, G., et al. Annual Review of Public Health. 2016. Assessing the Exposome with External Measures: Commentary on the State of the Science and Research Recommendations.
31. Dennis, KK., Marder, E., Balshaw, DM., Cui, Y., Lynes, MA., Patti, GJ., et al. Environmental health perspectives. 2016. Biomonitoring in the Era of the Exposome.
32. Dennis KK, Auerbach SS, Balshaw DM, Cui Y, Fallin MD, Smith MT, et al. The importance of the biological impact of exposure to the concept of the exposome. Environ Health Perspect. 2016; 124:1504–10. [PubMed: 27258438]
33. Rappaport SM, Smith MT. Epidemiology. Environment and disease risks. Science. 2010; 330:460–1. [PubMed: 20966241]
34. Wild CP. The exposome: from concept to utility. Int J Epidemiol. 2012; 41:24–32. [PubMed: 22296988]
35. National Institutes of Health, National Institute of Environmental Health Sciences. Children's Health Exposure Analysis Resource (CHEAR). 2016. 10/11/16]; Available from: https://www.niehs.nih.gov/research/supported/exposure/chear/
36. Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, et al. The human early-life exposome (HELIX): project rationale and design. Environmental health perspectives. 2014; 122:535–44. [PubMed: 24610234]
37. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011; 473:174–80. [PubMed: 21508958]
38. Backhed F. 99th Dahlem conference on infection, inflammation and chronic inflammatory disorders: the normal gut microbiota in health and disease. Clin Exp Immunol. 2010; 160:80–4. [PubMed: 20415855]
39. Collado MC, Rautava S, Isolauri E, Salminen S. Gut microbiota: a source of novel tools to reduce the risk of human disease? Pediatr Res. 2015; 77:182–8. [PubMed: 25335085]
40. Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome Med. 2011; 3:14. [PubMed: 21392406]
41. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A. 2011; 108(Suppl 1): 4578–85. [PubMed: 20668239]
42. Lee YK, Mazmanian SK. Has the microbiota played a critical role in the evolution of the adaptive immune system? Science. 2010; 330:1768–73. [PubMed: 21205662]
43. Lynch SV, Boushey HA. The microbiome and development of allergic disease. Curr Opin Allergy Clin Immunol. 2016; 16:165–71. [PubMed: 26885707]
44. Macfarlane GT, Macfarlane LE. Acquisition, evolution and maintenance of the normal gut microbiota. Dig Dis. 2009; 27(Suppl 1):90–8. [PubMed: 20203503]
45. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016; 8:51. [PubMed: 27122046]
46. Belkaid Y, Tamoutounour S. The influence of skin microorganisms on cutaneous immunity. Nat Rev Immunol. 2016; 16:353–66. [PubMed: 27231051]
47. Cundell AM. Microbial ecology of the human skin. Microb Ecol. 2016
48. Grassl N, Kulak NA, Pichler G, Geyer PE, Jung J, Schubert S, et al. Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. Genome Med. 2016; 8:44. [PubMed: 27102203]

49. Mammen MJ, Sethi S. COPD and the microbiome. Respirology. 2016; 21:590–9. [PubMed: 26852737]
50. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics. 2016
51. Shaw JG, Vaughan A, Dent AG, O'Hare PE, Goh F, Bowman RV, et al. Biomarkers of progression of chronic obstructive pulmonary disease (COPD). J Thorac Dis. 2014; 6:1532–47. [PubMed: 25478195]
52. Robinson CK, Brotman RM, Ravel J. Intricacies of assessing the human microbiome in epidemiologic studies. Ann Epidemiol. 2016; 26:311–21. [PubMed: 27180112]
53. Mai V, Prosperi M, Yaghjyan L. Moving microbiota research toward establishing causal associations that represent viable targets for effective public health interventions. Ann Epidemiol. 2016; 26:306–10. [PubMed: 27133768]
54. Debelius JW, Vazquez-Baeza Y, McDonald D, Xu Z, Wolfe E, Knight R. Turning participatory microbiome research into usable data: lessons from the American Gut Project. J Microbiol Biol Educ. 2016; 17:46–50. [PubMed: 27047589]
55. Almeida M, Pop M, Le Chatelier E, Prifti E, Pons N, Ghozlane A, et al. Capturing the most wanted taxa through cross-sample correlations. ISME J. 2016
56. Sinha R, Abnet CC, White O, Knight R, Huttenhower C. The microbiome quality control project: baseline study design and future directions. Genome Biol. 2015; 16:276. [PubMed: 26653756]
57. Morgan XC, Huttenhower C. Chapter 12: Human microbiome analysis. PLoS Comput Biol. 2012; 8:e1002808. [PubMed: 23300406]
58. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012; 22:292–8. [PubMed: 22009990]
59. Athersuch TJ. The role of metabolomics in characterizing the human exposome. Bioanalysis. 2012; 4:2207–12. [PubMed: 23046263]
60. Tzoulaki I, Ebbels TM, Valdes A, Elliott P, Ioannidis JP. Design and analysis of metabolomics studies in epidemiologic research: a primer on-omic technologies. Am J Epidemiol. 2014; 180:129–39. [PubMed: 24966222]
61. Tang WH, Wang Z, Kennedy DJ, Wu Y, Buffa JA, Agatisa-Boyle B, et al. Gut microbiota-dependent trimethylamine N-oxide (TMAO) pathway contributes to both development of renal insufficiency and mortality risk in chronic kidney disease. Circ Res. 2015; 116:448–55. [PubMed: 25599331]
62. Tang WH, Wang Z, Levison BS, Koeth RA, Britt EB, Fu X, et al. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. N Engl J Med. 2013; 368:1575–84. [PubMed: 23614584]
63. Lee IM, Shiroma EJ. Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges. Br J Sports Med. 2014; 48:197–201. [PubMed: 24297837]
64. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med. 2014; 48:1019–23. [PubMed: 24782483]
65. Buman MP, Hu F, Newman E, Smeaton AF, Epstein DR. Behavioral periodicity detection from 24 h wrist accelerometry and associations with cardiometabolic risk and health-related quality of life. Biomed Res Int. 2016; 2016:4856506. [PubMed: 26942195]
66. Kate RJ, Swartz AM, Welch WA, Strath SJ. Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data. Physiol Meas. 2016; 37:360–79. [PubMed: 26862679]
67. Chomistek AK, Shiroma EJ, Lee IM. The relationship between time of day of physical activity and obesity in older women. J Phys Act Health. 2016; 13:416–8. [PubMed: 26445277]
68. Nethery E, Mallach G, Rainham D, Goldberg MS, Wheeler AJ. Using Global Positioning Systems (GPS) and temperature data to generate time-activity classifications for estimating personal exposure in air monitoring studies: an automated method. Environ Health. 2014; 13:33. [PubMed: 24885722]

69. Stewart OT, Moudon AV, Fesinmeyer MD, Zhou C, Saelens BE. The association between park visitation and physical activity measured with accelerometer, GPS, and travel diary. Health Place. 2016; 38:82–8. [PubMed: 26798965]
70. Shearer C, Rainham D, Blanchard C, Dummer T, Lyons R, Kirk S. Measuring food availability and accessibility among adolescents: Moving beyond the neighbourhood boundary. Soc Sci Med. 2015; 133:322–30. [PubMed: 25619139]
71. Evanoff BA, Gustavsson P, Hogstedt C. Mortality and incidence of cancer in a cohort of Swedish chimney sweeps: an extended follow up study. Br J Ind Med. 1993; 50:450–9. [PubMed: 8507598]
72. Alavanja MC, Samanic C, Dosemeci M, Lubin J, Tarone R, Lynch CF, et al. Use of agricultural pesticides and prostate cancer risk in the Agricultural Health Study cohort. Am J Epidemiol. 2003; 157:800–14. [PubMed: 12727674]
73. Dosemeci M, Alavanja MC, Rowland AS, Mage D, Zahm SH, Rothman N, et al. A quantitative approach for estimating exposure to pesticides in the Agricultural Health Study. Ann Occup Hyg. 2002; 46:245–60. [PubMed: 12074034]
74. Cochran RC, Driver JH. Estimating human exposure: improving accuracy with chemical markers. Prog Mol Biol Transl Sci. 2012; 112:11–29. [PubMed: 22974736]
75. Thompson FE, Dixit-Joshi S, Potischman N, Dodd KW, Kirkpatrick SI, Kushi LH, et al. Comparison of Interviewer-Administered and Automated Self-Administered 24-Hour Dietary Recalls in 3 Diverse Integrated Health Systems. Am J Epidemiol. 2015; 181:970–8. [PubMed: 25964261]
76. Kirkpatrick SI, Subar AF, Douglass D, Zimmerman TP, Thompson FE, Kahle LL, et al. Performance of the Automated Self-Administered 24-hour Recall relative to a measure of true intakes and to an interviewer-administered 24-h recall. The American journal of clinical nutrition. 2014; 100:233–40. [PubMed: 24787491]
77. National Institutes of Health, National Cancer Institute. Automated Self-Administered 24-Hour (ASA24®) Dietary Assessment Tool. 2013. 10/11/16]; Available from: http://epi.grants.cancer.gov/asa24/
78. Zhu F, Bosch M, Woo I, Kim S, Boushey CJ, Ebert DS, et al. The Use of Mobile Devices in Aiding Dietary Assessment and Evaluation. IEEE journal of selected topics in signal processing. 2010; 4:756–66. [PubMed: 20862266]
79. Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional cohort studies. Journal of the National Cancer Institute. 2011; 103:1086–92. [PubMed: 21653922]
80. Patel CJ, Ioannidis JP. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. J Epidemiol Community Health. 2014; 68:1096–100. [PubMed: 24923805]
81. Gauderman WJ, Mukheerjee B, Aschard H, Hsu L, Lewinger JP, Patel CJ, et al. Update on the State of the Science for Analytical Methods. Am J Epidemiol. 2017 in press.
82. Choi JOM, AJ. Estimating the Causal Effect of Treatment in Observational Studies with Survival Time Endpoints and Unmeasured Confounding. 2014:1893–907.
83. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Med. 2007; 4:e352. [PubMed: 18076282]
84. Patel CJ, Manrai AK. Development of exposome globes to map out environment-wide associations. Pac Symp Biocomput. 2015
85. Patel CJ, Manrai AK. Development of exposome correlation globes to map out environment-wide associations. Pac Symp Biocomput. 2015:231–42. [PubMed: 25592584]
86. Patel CJ. Analytic Complexity and Challenges in Identifying Mixtures of Exposures Associated with Phenotypes in the Exposome Era. Current epidemiology reports. 2017; 4:22–30. [PubMed: 28251040]
87. Patel CJ, Ioannidis JP. Studying the elusive environment in large scale. JAMA. 2014; 311:2173–4. [PubMed: 24893084]

88. Manrai AK, Cui Y, Bushel PR, Hall M, Karakitsios S, Mattingly CJ, et al. Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health. Annu Rev Public Health. 2017; 38:279–94. [PubMed: 28068484]

89. Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, et al. The PhenX Toolkit: get the most from your measures. Am J Epidemiol. 2011; 174:253–60. [PubMed: 21749974]

90. Patel CJ. Analytical Complexity in Detection of Gene Variant-by-Environment Exposure Interactions in High-Throughput Genomic and Exposomic Research. Current environmental health reports. 2016; 3:64–72. [PubMed: 26809563]

91. Patel CJ, Chen R, Kodama K, Ioannidis JP, Butte AJ. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. Hum Genet. 2013; 132:495–508. [PubMed: 23334806]

92. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B, Statistical Methodology. 1995; 57:289–300.

93. Ritchie MD, Davis JR, Aschard H, Battle A, Conti D, Du M, et al. Incorporation of Biological Knowledge into the Study of GxE. Am J Epidemiol. 2017 in press.

94. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015; 348:660–5. [PubMed: 25954002]

95. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. Am J Epidemiol. 2009; 169:227–30. discussion 34–5. [PubMed: 19022826]

96. Thomas D. Gene-environment-wide association studies: emerging approaches. Nature reviews Genetics. 2010; 11:259–72.

97. Li S, Mukherjee B, Taylor JM, Rice KM, Wen X, Rice JD, et al. The role of environmental heterogeneity in meta-analysis of gene-environment interactions with quantitative traits. Genet Epidemiol. 2014; 38:416–29. [PubMed: 24801060]

98. Du M, Zhang X, Hoffmeister M, Schoen RE, Baron JA, Berndt SI, et al. No evidence of gene-calcium interactions from genome-wide analysis of colorectal cancer risk. Cancer Epidemiol Biomarkers Prev. 2014; 23:2971–6. [PubMed: 25192705]

99. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005; 2:e124. [PubMed: 16060722]

100. Ioannidis JP. Exposure-wide epidemiology: revisiting Bradford Hill. Stat Med. 2016; 35:1749–62. [PubMed: 26646432]

101. Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology. 2011; 22:450–6. [PubMed: 21490505]

102. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS One. 2010; 5:e10746. [PubMed: 20505766]

103. Xiao Q, Moore SC, Boca SM, Matthews CE, Rothman N, Stolzenberg-Solomon RZ, et al. Sources of variability in metabolite measurements from urinary samples. PLoS One. 2014; 9:e95749. [PubMed: 24788433]

104. Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. Am J Epidemiol. 1997; 146:596–604. [PubMed: 9326439]

105. Garcia-Closas M, Thompson WD, Robins JM. Differential misclassification and the assessment of gene-environment interactions in case-control studies. Am J Epidemiol. 1998; 147:426–33. [PubMed: 9525528]

106. Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. Cancer Epidemiol Biomarkers Prev. 1999; 8:1043–50. [PubMed: 10613335]

107. Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? Int J Epidemiol. 2003; 32:51–7. [PubMed: 12690008]

108. Wong MY, Day NE, Luan JA, Wareham NJ. Estimation of magnitude in gene-environment interactions in the presence of measurement error. Stat Med. 2004; 23:987–98. [PubMed: 15027084]
109. Cheng KF. Analysis of case-only studies accounting for genotyping error. Ann Hum Genet. 2007; 71:238–48. [PubMed: 16984488]
110. Zhang L, Mukherjee B, Ghosh M, Gruber S, Moreno V. Accounting for error due to misclassification of exposures in case–control studies of gene–environment interaction. Stat Med. 2008; 27:2756–83. [PubMed: 17879261]
111. Lindstrom S, Yen YC, Spiegelman D, Kraft P. The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. Hum Hered. 2009; 68:171–81. [PubMed: 19521099]
112. Boonstra PS, Mukherjee B, Gruber SB, Ahn J, Schmit SL, Chatterjee N. Tests for gene-environment interactions and joint effects with exposure misclassification. Am J Epidemiol. 2016; 183:237–47. [PubMed: 26755675]
113. Lobach I, Carroll RJ, Spinka C, Gail MH, Chatterjee N. Haplotype-based regression analysis and inference of case-control studies with unphased genotypes and measurement errors in environmental exposures. Biometrics. 2008; 64:673–84. [PubMed: 18047538]
114. Vanderweele TJ. Inference for additive interaction under exposure misclassification. Biometrika. 2012; 99:502–8. [PubMed: 23843668]
115. Stenzel SL, Ahn J, Boonstra PS, Gruber SB, Mukherjee B. The impact of exposure-biased sampling designs on detection of gene-environment interactions in case-control studies with potential exposure misclassification. Eur J Epidemiol. 2015; 30:413–23. [PubMed: 24894824]
116. 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; 175:203–7. discussion 8–9. [PubMed: 22199029]
117. Gauderman WJ, Mukheerjee B, Aschard H, Hsu L, Lewinger JP, Patel CJ, et al. Update on the State of the Science for Analytical Methods. American Journal of Epidemiology. 2016 submitted manuscript.
118. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics. 2010; 26:445–55. [PubMed: 20053841]
119. Hunter DJ. Gene-environment interactions in human diseases. Nature reviews Genetics. 2005; 6:287–98.
120. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Hum Hered. 2007; 63:111–9. [PubMed: 17283440]
121. Qi Q, Chu AY, Kang JH, Jensen MK, Curhan GC, Pasquale LR, et al. Sugar-sweetened beverages and genetic risk of obesity. N Engl J Med. 2012; 367:1387–96. [PubMed: 22998338]
122. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med. 2008; 359:2208–19. [PubMed: 19020323]
123. Park SK, Tao Y, Meeker JD, Harlow SD, Mukherjee B. Environmental risk score as a new tool to examine multi-pollutants in epidemiologic research: an example from the NHANES study using serum lipid levels. PLoS One. 2014; 9:e98632. [PubMed: 24901996]
124. Patel CJ, Cullen MR, Ioannidis JP, Butte AJ. Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels. Int J Epidemiol. 2012; 41:828–43. [PubMed: 22421054]
125. Patel CJ, Rehkopf DH, Leppert JT, Bortz WM, Cullen MR, Chertow G, et al. Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States National Health and Nutrition Examination Survey. Int J Epidemiol. 2013; 42:1795–810. [PubMed: 24345851]
126. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013; 9:e1003348. [PubMed: 23555274]
127. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet. 2001; 29:306–9. [PubMed: 11600885]

128. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature. 2008; 452:638–42. [PubMed: 18385739]
129. Bierut LJ, Stitzel JA, Wang JC, Hinrichs AL, Grucza RA, Xuei X, et al. Variants in nicotinic receptors and risk for nicotine dependence. Am J Psychiatry. 2008; 165:1163–71. [PubMed: 18519524]
130. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. 2008; 40:616–22. [PubMed: 18385676]
131. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008; 452:633–7. [PubMed: 18385738]
132. Chen LS, Baker TB, Piper ME, Breslau N, Cannon DS, Doheny KF, et al. Interplay of genetic risk factors (CHRNA5-CHRNA3-CHRNB4) and cessation treatments in smoking cessation success. Am J Psychiatry. 2012; 169:735–42. [PubMed: 22648373]
133. Chen LS, Horton A, Bierut L. Pathways to precision medicine in smoking cessation treatments. Neuroscience letters. 2016
134. Chen LS, Hung RJ, Baker T, Horton A, Culverhouse R, Saccone N, et al. CHRNA5 risk variant predicts delayed smoking cessation and earlier lung cancer diagnosis--a meta-analysis. J Natl Cancer Inst. 2015:107.
135. Holmes E, Loo RL, Stamler J, Bictash M, Yap IK, Chan Q, et al. Human metabolic phenotype diversity and its association with diet and blood pressure. Nature. 2008; 453:396–400. [PubMed: 18425110]
136. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, et al. Metabolite profiles and the risk of developing diabetes. Nat Med. 2011; 17:448–53. [PubMed: 21423183]
137. Chen, J., Marshall, SJ., Wang, L., Godbole, S., Legge, A., Doherty, A., et al. Using the SenseCam as an objective tool for evaluating eating patterns. Proceedings of the 4th International SenseCam & Pervasive Imaging Conference; 2013; ACM; p. 34-41.
138. Ellis, K., Godbole, S., Chen, J., Marshall, S., Lanckriet, G., Kerr, J. Physical activity recognition in free-living from body-worn sensors. Proceedings of the 4th International SenseCam & Pervasive Imaging Conference; 2013; ACM; p. 88-9.
139. Marinac, C., Merchant, G., Godbole, S., Chen, J., Kerr, J., Clark, B., et al. The feasibility of using SenseCams to measure the type and context of daily sedentary behaviors. Proceedings of the 4th International SenseCam & Pervasive Imaging Conference; 2013; ACM; p. 42-9.
140. Lam, MS., Godbole, S., Chen, J., Oliver, M., Badland, H., Marshall, SJ., et al. Measuring time spent outdoors using a wearable camera and GPS. Proceedings of the 4th International SenseCam & Pervasive Imaging Conference; 2013; ACM; p. 1-7.
141. Subar AF, Thompson FE, Kipnis V, Midthune D, Hurwitz P, McNutt S, et al. Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. Am J Epidemiol. 2001; 154:1089–99. [PubMed: 11744511]
142. National Institutes of Health, National Cancer Institute. Dietary Assessment Primer. Oct 11. 2016]; Available from: https://dietassessmentprimer.cancer.gov/
143. Kerr J, Patterson RE, Ellis K, Godbole S, Johnson E, Lanckriet G, et al. Objective Assessment of Physical Activity: Classifiers for Public Health. Med Sci Sports Exerc. 2016; 48:951–7. [PubMed: 27089222]
144. Meseck K, Jankowska MM, Schipperijn J, Natarajan L, Godbole S, Carlson J, et al. Is missing geographic positioning system data in accelerometry studies a problem, and is imputation the solution? Geospat Health. 2016; 11:403. [PubMed: 27245796]
145. Jerrett M, Burnett RT, Ma R, Pope CA 3rd, Krewski D, Newbold KB, et al. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology. 2005; 16:727–36. [PubMed: 16222161]
146. Jankowska MM, Schipperijn J, Kerr J. A framework for using GPS data in physical activity and sedentary behavior studies. Exerc Sport Sci Rev. 2015; 43:48–56. [PubMed: 25390297]
147. O'Connell SG, Kincl LD, Anderson KA. Silicone wristbands as personal passive samplers. Environ Sci Technol. 2014; 48:3327–35. [PubMed: 24548134]

Table 1

Examples of traditional and emerging environmental and behavioral measurements on the individual level (i) and ecological level (e).

Description	Examples references	Type (sensor or bioassay; external or internal?); How implemented/disseminated (tissue sample, monitoring device)	Rough cost/ participant (O: order notation)	# of Variables	Sources of error (e.g., measurement)
Microbiome (i)	Robinson CK, et al. (52)	Sequencing of samples (e.g., feces, saliva)	O($100)	O(1000)	Diverse ‘omics modalities, diverse sample collection methods
Targeted mass spectrometry and biomarkers (i)	Holmes et al(135), Wang et al. (136)	Assay of human tissue (e.g., serum, urine, tissue-specific)	O($100–1000)	O(10–100)	Technical variation; sample origin and collection
Untargeted mass spectrometry/metabolomics (i)	Tzoulaki I, et al.(60)	Assay of human tissue (e.g., serum, urine, tissue-specific)	O($100–1000)	O(1000)	Technical variation; sample origin and collection
Context & behavior assessment (i)	Chen J, et al. (137); Ellis K, et al. (138); Marinac C, et al. (139); Lam MS, et al. (140)	SenseCam camera; iPhone research kit apps	<$100	O(10–100)	Device use and attrition; recall
Dietary intake assessment(i)	Subar et al (141) Thompson et al. (75, 142)	Self-administered auto-coded mobile and/or web-based 24-hour recalls, food records and food frequency questionnaires	Some freely available, others < $100	O(100)	Day-to-day random error and systematic error or bias
Physical activity assessment (i)	Kerr J, et al. (143); Meseck K, et al. (144)	Accelerometer	<$100	O(10–100)	Wear time
Occupational Exposures assessments	Cochran RC, Driver JH (74)	Questionnaire supported by algorithms to infer exposures	<$100	O(1000)	Participant recall, imprecision estimating exposures
Air pollution monitoring (e)	Jerrett, M, et al. (145)	Sensor	$100–1000	O(10–100)	Imprecision estimating internal exposure
Global Positioning System (e)	Jankowska MM et al. (146)	Sensor	<$100	O(10–100)	Location error
Individual sensors (i)	O’Connell et al.(147)	Sensor	$100–1000	O(10–100)	Measurement error; reporting bias; wear time

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2018 September 01.