PERSPECTIVE

Research on Gene–Environment Interplay in the Era of “Big Data”

ABSTRACT. Successful identification of genetic risk factors in genome-wide association studies typically has depended on meta-analyses combining data from large numbers of studies involving tens or hundreds of thousands of participants. This poses a challenge for research on Gene $\times$ Environment interaction ($G \times E$) effects, where characterization of environmental exposures is quite limited in most studies and often varies idiosyncratically between studies. Yet the importance of environmental exposures in the etiology of many disorders—and especially alcohol, tobacco, and drug use disorders—is undeniable. We discuss the potential

for “big-data” approaches (e.g., aggregating data from state databases) to generate consistent measures of neighborhood environment across multiple studies, requiring only information about residential address (or ideally residential history) to make progress in $G \times E$ analyses. Big-data approaches may also help address limits to the generalizability of existing research literature, such as those that arise because of the limited numbers of severely alcohol-dependent mothers represented in prospective research studies. (*J. Stud. Alcohol Drugs*, 77, 681–683, 2016)

AS WITH OTHER COMPLEX (NON-MENDELIAN) medical phenotypes, the identification of specific genetic variants that increase risk of a given substance use disorder has proved challenging, but continuing progress is to be anticipated. We have moved beyond the era of small sample candidate gene studies, arguably an era of random-walk science, with multiple independent teams of investigators pursuing their favorite candidate genes (albeit frequently informed by data from animal models). In hindsight, for addictions as for other complex medical phenotypes, the primary product of that research was a massive accumulation of false-positive publications (Ioannidis et al., 2011)—with rare exceptions, such as genes that influence drug metabolism (e.g., MacGregor et al., 2009).

From genome-wide association studies of complex phenotypes, we now know that there may be many hundreds of genetic variants making slight contributions to risk differences, with an accumulation of perhaps many hundred thousand cases through meta-analyses of many data sets needed for discovery of such variants (e.g., for body mass index [BMI], Locke et al., 2015, using data from several hundred thousand research participants, found 97 variants that are genome-wide significant and that together account for 2.7% of the variance in BMI). As whole genome sequencing studies of nonpsychiatric medical conditions become routine, accumulation of the associated medical record data on alcohol and tobacco use disorders—which have high rates of comorbidity with many medical conditions—will ensure continuing progress in gene discovery for alcohol and tobacco use disorders, requiring only investment in data aggregation. For illicit drug use disorders, less likely to be noted in the medical record,

progress (and the return on the necessary high capital investment) will be more uncertain.

Substance use disorders provide perhaps the most compelling examples of gene–environment interplay in psychiatry, with the joint effects of genetic risk and environmental risk exposures likely having become increasingly important as the prevalence of alcohol and drug use disorders in women (still most often the primary caretakers of children) has increased. The alcohol and drug abuse fields have accumulated many rich data sets that provide information about environmental risk exposures, albeit information that has as yet been imperfectly mined. Yet how can we minimize the risk of random-walk science? Ioannidis (2005) has shown that a high rate of false-positive biomedical publications is to be anticipated whenever the a priori probability of a true positive finding is low (always true of innovative science), statistical power is marginal, multiple testing is likely because of a large universe of potential variables that might be tested or multiple ways of operationalizing a construct or analyzing data are possible, and multiple teams of investigators are working independently on a problem (with only teams obtaining positive results likely to publish). These are exactly the conditions that currently exist for analyses of Gene $\times$ Environment interaction and that existed previously for candidate gene studies.

An urgent challenge, therefore, is how to advance understanding of the interplay of genetic and environmental risk mechanisms while minimizing the risk of random-walk science. We can learn much from “big-data” approaches that, for commercial goals, may use many tens or hundreds of millions of records. The addiction research community can

do much more to link research data with individual-level data from population databases such as, in the U.S. context, vital records (e.g., birth record data) and driver's license data (to which access for secondary data analysis is allowed by U.S. federal law [18 U.S. Code § 2721] [Drivers Privacy Protection Act, 1994], although individual states may have more restrictive policies). Such population databases typically have a very standard structure (e.g., most state birth record databases follow a standard model, albeit with idiosyncratic omissions for some states), allowing data to be aggregated for multiple states and linked to existing research data sets in a meta-analytic framework.

Many steps will be needed in such data-linkage efforts:

(a) *Using research data to cross-validate information in a state database, and vice versa, to better understand error rates and systematic biases (e.g., maternal age may have been misstated in the birth record for a mother still a minor).*

(b) *Using research data from prospective cohort studies (and in particular the residential history information that is obtained) to cross-validate interval-censored data about residential mobility patterns that may be derived from state databases (e.g., updated parental and youth addresses obtained via state identification/driver's license data at each renewal or address change).* The goal here is to be able to model mobility patterns of an entire population, not just of research participants, moving beyond a static, cross-sectional approach that views current environment as the major determinant of current phenotype to a more developmentally informed approach that recognizes movement of individuals between neighborhoods.

(c) *Aggregating data from state databases to provide neighborhood risk indices that supplement standard census, alcohol outlet, taxation, and similar measures and that allow for changing neighborhood risk factors over time.* For example, rates of driving under the influence (DUI) arrests, births to teen mothers, maternal smoking during pregnancy, and overweight/obesity (the latter perhaps well approximated from data obtained from first, but not subsequent, application for state identification or driver's license) may all prove useful risk indicators.

(d) *Using state data to address limits to the external validity (generalizability) of data from a research study.* From state birth record data, for example, we can construct synthetic (in silico) cohorts whose sociodemographic background and outcomes (e.g., maternal characteristics; DUI convictions, reproduction, mobility patterns) are contrasted with those of more specialized research samples (e.g., from a high-risk prospective study). Because severe alcoholism in women is relatively rare, most research studies will provide information only about mothers with milder alcohol use disorder. Yet by merging variables from vital records data with DUI data, we can begin to address in women with recurrent drunk-driving arrests (i.e., those who typically having more severe problems) outcomes such as own and child mortality,

reproductive timing, separation from reproductive partner, and downward social mobility. This will allow us to address the generalizability of the findings observed in less severely affected research participants.

How do we proceed? To consider just one example, parental separation is a strong predictor of early-onset substance use, even after adjustment for parental substance use disorder histories (e.g., Waldron et al., 2014), and risk of parental separation is highly predictable from variables such as maternal age, educational level, and marital status at childbirth—all of which may be derived from the birth record. Do neighborhoods with increased rates of family separation predict increased rates of early-onset substance use, even after statistical control for other neighborhood risk indices and for individual and family-level variables? Do such neighborhood characteristics exacerbate the effects of pre-existing genetic vulnerability?

Risk probabilities estimated via statistical methods such as propensity score analysis (Imbens & Rubin, 2015), using data from hundreds of thousands, can be aggregated as a continuous neighborhood-level predictor to determine the ability to improve prediction of substance use outcomes in research data, both in models incorporating only neighborhood-level predictors and in models incorporating family-level, individual, and genetic data from research studies, including tests for interaction effects. By working at the level of identified individuals (mothers) in state databases, we can allow for the change of residential neighborhood composition in the population through time as well as for the geographic mobility of our research participants. Such an approach uses some strong simplifications (e.g., will ignore mothers who move into a state after the birth of a child, until such time that the data can be aggregated across all states). However, given the consistency of database structures across U.S. states, the approach is scalable to allow meta- or mega-analysis combining data from many existing research studies throughout the United States, a necessary first step to advance research on gene-environment interplay for some simple and consistently defined environmental exposure measures.

Finally, what are the ethical and human-subject implications of such big-data approaches? The Common Rule—the U.S. regulatory framework for Human Subjects research (U.S. Department of Health and Human Services, 2009)—allows an institutional review board to waive the requirement to obtain informed consent (45 C.F.R. 46.116(d)), a waiver that is a requirement for such research. Protection of confidentiality can be maximized by generating a “third numeric code” for every unique individual across all databases that is distinct from both the alphanumeric codes associated with state databases (e.g., state driver's license number, which of course cannot directly be appended to research data because it is an identifier) and the alphanumeric subject identifiers associated with research databases. It requires maintaining

a clear separation of roles among (a) those who work with traditional subject identifiers (name, date of birth) to generate the unique third numeric code and create the database table that cross-links the three identifier sets (who will have no access to research data); (b) those who work with research data who may also receive aggregate neighborhood variables for research participants but who will have access to neither traditional subject identifiers nor the codes that link to the state databases; and (c) those few key individuals with access privileges to join at the individual level who can select variables from research and state databases. In this way, those engaged in data analysis are working with de-identified data, and those working with traditional subject identifiers have no access to research variables. Use of database systems with strong audit capabilities and stringent access control also seems like a requirement to ensure against unauthorized database use. Such capabilities are, of course, already an expectation for clinical data covered by the Health Insurance Portability and Accountability Act (HIPAA). With such safeguards, risks of confidentiality breach can be minimized while maximizing the potential to advance Genome $\times$ Environment interaction research.

ANDREW C. HEATH, D.PHIL.^{a,b}
aheath@wustl.edu

CHRISTINA N. LESSOV-SCHLAGGAR, PH.D.^{a,b}

MIN LIAN, M.D., PH.D.^{a,c}

RUTH MILLER, PH.D.^{a,b}

ALEXIS E. DUNCAN, PH.D.^{a,d}

PAMELA A. F. MADDEN, PH.D.^{a,b}

^aMidwest Alcoholism Research Center

^bDepartment of Psychiatry,
Washington University School of Medicine,
St. Louis, Missouri

^cDepartment of Medicine,
Washington University School of Medicine,
St. Louis, Missouri

^dGeorge Warren Brown School of Social Work,
Washington University in St. Louis,
St. Louis, Missouri

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

References

- Drivers Privacy Protection Act, 18 U.S.C. § 2721 (1994). Retrieved from <http://www.accessreports.com/statutes/DPPA1.htm>
- Imbens, G. W., & Rubin, D. R. (2015). *Causal inference for statistics, social and biomedical sciences: An introduction*. Cambridge, England: Cambridge University Press.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. doi:10.1371/journal.pmed.0020124
- Ioannidis, J. P., Tarone, R., & McLaughlin, J. K. (2011). The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*, 22, 450–456. doi:10.1097/EDE.0b013e31821b506e
- Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., . . . Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518, 197–206. doi:10.1038/nature14177
- Macgregor, S., Lind, P. A., Bucholz, K. K., Hansell, N. K., Madden, P. A. F., Richter, M. M., . . . Whitfield, J. B. (2009). Associations of *ADH* and *ALDH2* gene variation with self report alcohol reactions, consumption and dependence: An integrated analysis. *Human Molecular Genetics*, 18, 580–593. doi:10.1093/hmg/ddn372
- U.S. Department of Health and Human Services. (2009). *Code of Federal Regulations: Title 45 Public Welfare. Part 46: Protection of Human Subjects*. Retrieved from <http://www.hhs.gov/ohrp/sites/default/files/ohrp/policy/ohrpregulations.pdf>
- Waldron, M., Vaughan, E. L., Bucholz, K. K., Lynskey, M. T., Sartor, C. E., Duncan, A. E., . . . Heath, A. C. (2014). Risks for early substance involvement associated with parental alcoholism and parental separation in an adolescent female cohort. *Drug and Alcohol Dependence*, 138, 130–136. doi:10.1016/j.drugalcdep.2014.02.020