Genetic Self Knowledge and the Future of Epidemiologic Confounding

Tyler Vander Weele[1,*]

Prior work has considered how our genetic knowledge might allow for personalized medicine. This commentary explores the reverse question of what personalized genetic medicine might do to our research process, not only in genetics, but in epidemiology more generally.

Introduction

A number of recent articles have considered the extent to which our genetic knowledge and research can or eventually will allow for personalized genetic medicine.[1–6] Although for many diseases our capacity to utilize knowledge of an individual's genome to predict risk is still limited,[7–10] research and technological development continue to progress at a rapid pace.[6,8] Moreover, risks tests are already available for certain diseases,[5,11,12] and direct-to-consumer profiling is a present reality.[13]

The potential of individual genomic information for personalized medicine will likely continue to be the topic of intense discussion, and it will be important to reflect upon the implications for medicine of advances made in genetic science.[6] In this commentary, however, I would like to briefly consider the reverse question: not what genomic research will do for personalized medicine, but rather, what personalized genetic medicine may do to our research process, not only in genetics but in epidemiology more generally. Stated simply, as our understanding of genetic risk advances and as individuals acquire knowledge of their own genomes, health behaviors are likely to change and an individual's knowledge of his or her genetic risk may start confounding the relationship between disease and environmental exposures when no confounding was previously present. This commentary offers a few reflections on the implications of genetic self knowledge for confounding and ascertainment bias, for what study designs and analytic techniques may be appropriate in research, and for how we might prepare for the altered research landscape that may result from personalized genetic medicine.

Implications of Genetic Self Knowledge for Research

As genotyping becomes increasingly affordable and as our understanding of the genetic basis of disease progresses, the demand for personal genetic information will likely also increase. If, eventually, it does become possible to accurately assess an individual's genetic risk, knowledge of such risk may change individuals' behaviors and actions. High-risk individuals may aggressively seek to avoid behavioral or environmental exposures that increase risk yet further. It is possible that such information may eventually be used in constructing effective prevention programs.[6] When substantial gene-environment interactions are present and known, this may even further increase motivation to avoid or eliminate certain environmental exposures. Such alterations, if they occur, could undoubtedly be counted as a success for genetic preventive medicine. However, an unintended consequence of changes in behavior that result from genetic self knowledge would be that an individual's genetic risk factors would then suddenly serve as confounding factors for the relationship between behavioral and environmental exposures and disease where none existed before. With personal genetic knowledge, the genetic risk factors for a particular disease (mediated by an individual's knowledge of them) would themselves affect both the likelihood of exposure to the environmental risk factors and also the likelihood of developing the disease. This would arise even if there were no biological or natural link between the genetic factor and the environmental factor to begin with; rather, the confounding would arise through an individual's knowledge of their own genetic risk and by an alteration of behavior resulting from such knowledge. A study that did not control for such confounding by genetic self knowledge could end up with underestimates of the effect of the environmental factor on the disease, given that individuals with the highest genetic risk may be those that have intentionally taken action to ensure the lowest possible level of environmental exposure. If individuals with a family history of disease are also more likely to make use of genetic testing, this would further strengthen possible confounding.

Advances in our understanding of the genetic basis of disease already create some possibility of this occurring in the near future. For example, it was recently noted that by making use of 12 recently discovered variants associated with the risk of myocardial infarction (MI) it is now possible to identify 10% of populations of

[1]Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA
*Correspondence: tvanderw@hsph.harvard.edu
DOI 10.1016/j.ajhg.2010.07.006.

European descent who have a 1.6× elevated risk of MI on account of their genetic profiles.[6] If this knowledge is generally available to the individuals with elevated risk, this may result in a change of health behaviors concerning diet, smoking, and exercise. Likewise, individuals with low genetic risk of MI may end up making less strenuous efforts at altering environmental and behavioral MI risk factors. The genetic factors for MI, or at least an individual's knowledge of them, then become confounding factors for the relationship between environmental exposures (e.g., diet, smoking, exercise) and MI. A study that did not control for these genetic factors (or for "genetic self knowledge") and examined the associations between environmental factors and MI could end up with biased estimates. For example, a risk ratio for MI of 1.3 comparing different diets could be completely eliminated by confounding from a genetic variant that increased MI risk by a factor of 1.6 and differed in prevalence by 60% between the groups with different diets; an actual risk ratio of 1.5 could be reduced by such confounding to approximately 1.1.[14]

Although the scope of this problem is likely to be small at present, its relevance may increase considerably with time as our knowledge of the genetic risk for common diseases advances and as personal genetic information becomes more widely available. Proposals have been made concerning the possibility of eventually using individually tailored lifelong programs of risk reduction as a future public health effort.[6] If this were to occur, the confounding structures introduced by such highly tailored lifelong programs, based on a potentially long list of genetic risk factors, could become quite complicated. Of course, it remains yet to be seen the extent to which genetic information will be of use in risk prediction and the extent to which behavior may change in light of knowledge of genetic risk.

One of the interesting features of the changes that are likely to take place in the knowledge that individuals have of their own genome is that the implications of this knowledge are relevant not simply for genetics research but also for more traditional epidemiologic analysis of environmental factors. Even if a chronic-disease epidemiologist is interested only in the effects of environmental exposures, genetic factors will come to serve as confounders in the study of such environmental factors. It may thus eventually become necessary to collect data on genetic risk factors even in studies in which the interest lies solely in assessing the effects of an environmental exposure, a point to which we will return below.

If personal genetic knowledge does begin to change behavioral and environmental exposures in this way, then the overall "effect" of genetic factors on various diseases is also likely to change. Genetic association studies conducted after such personal genetic knowledge is being used will capture not only the "biological" effects of various genetic risk factors, as is the case at present, but also the effects that these risk factors have on decisions about modifying environmental exposures to protect individuals from disease. Because these two effects will likely operate in opposite directions, associations between genetic variants and disease may end up being attenuated in future studies as individuals make use of genetic knowledge to make behavioral and lifestyle changes. Some of the significant findings may disappear even if the actual biology remains unchanged, and an awareness of this possibility will be important in not discarding in the future significant associations found before personal genetic knowledge becomes more widespread.

On the other hand, a recent web-based study using data from 23andMe makes clear that genetic self knowledge can also generate bias in the other direction when individuals self report their phenotype.[13] If phenotype is self reported, then individuals who know their genetic predictions for a particular trait may be more likely to self report the trait for which they are genetically predisposed. In analyses that consider sprinter versus long-distance runner as the phenotype, Eriksson et al.[13] found that responses differed considerably ($p < 10^{-63}$) between individuals who had or had not seen their genotypes, demonstrating that the degree of this form of ascertainment bias can be quite substantial.

Personal genetic knowledge has implications not only for introduction of confounding in epidemiologic research, for effect attenuation in genetic association studies, and for ascertainment bias for self-reported outcomes, but also for what study designs and analytic techniques may be appropriate. The case-only study design[15] has allowed for the analysis of gene-environment interaction in settings in which information is available only about the "cases" (i.e., diseased subjects), and this study design has also allowed for more powerful tests for interaction in a number of settings.[16] The study design, however, relies on an assumption of independence between the genetic and environmental factors. In some contexts this may be, though is not always,[15,17] a reasonable assumption. However, even if genetic and environmental factors are presently independent, if in the future personal genetic knowledge is used to make changes to an individual's environmental exposure, the assumption of independence will no longer be preserved. In samples in which genetic self knowledge is being used to change environmental exposure, the case-only design will no longer constitute a valid design (though if the genetic and environmental factors are negatively correlated, the case-only design may still yield conservative estimates of gene-environment interaction parameters[18,19]).

Likewise, certain family-based designs also rely on an assumption of independence of genetic and environmental factors,[20] and these too will be rendered inapplicable if personal genetic knowledge introduces correlation. Similarly, certain

statistical techniques used in the analysis of gene-environment interaction have also relied on an assumption of independence of genetic and environmental factors,[21–23] and these will likewise be rendered invalid if individuals use genetic self knowledge to alter their environmental exposures so that genetic and environmental factors become correlated. The implications of personal genetic knowledge thus not only simply concern the possibility of new sources of confounding and ascertainment bias but also extend to the validity of study designs and analytic techniques.

Preparing for Personal Genetics in Epidemiologic Research

Because of limited access to individual genomic information and our still very incomplete understanding of the genetic risk factors for common diseases, the problem of confounding from genetic self knowledge is probably of more relevance to the future than to the present. It will be helpful, however, to prepare for what may well become an increasingly widespread issue.

If confounding from such genetic self knowledge is present, it will in many instances be necessary to collect data on and control for the genetic confounding factors or to at least collect data on the knowledge that individuals have of this genetic information. Although, at present, genetic confounding, if operative, could jeopardize or render infeasible a great deal of current epidemiologic research because of the cost of genotyping entire study samples, this issue of prohibitive cost may become increasingly less problematic. The cost of genotyping has fallen considerably since the technology was first introduced, and the trend of declining costs is likely to continue. The possibility of obtaining genomic information in all epidemiologic studies, not for the purposes of assessing genetic association, but for the purposes of confounding control, may with time not be infeasible.

It should be noted that the possibility of genetic factors serving as confounders for the relationship between environmental exposures and various diseases can arise without personal knowledge of individual genetic variants. First, even without information on specific genetic variants, individuals often have knowledge of family history that may similarly alter behavior. However, for rare outcomes, the extent of the confounding bias that this generates may be small, and moreover, at least at present, it is easier and cheaper to control for family history than for specific genetic variants. Second, such genetic confounding can also occur if correlated genetic variants or the same genetic variant is a risk factor for both the exposure and the disease. For example, some recent findings indicate that certain genetic variants may be a common cause of both smoking behavior through nicotine dependence and lung cancer.[24–28] If this is indeed the case, this would bias effect estimates for smoking if not controlled for, although this bias would likely not be of sufficient magnitude to change qualitative conclusions.[29]

The smoking and lung cancer example suggests another way to move forward with rigorous research in the epidemiologic analysis of environmental and behavioral factors even if data is not available on potential genetic confounding. The possibility of an unobserved genetic factor affecting both smoking and lung cancer was proposed fairly early by Fisher.[30] Cornfield and colleagues[29] used associations between smoking and lung cancer from observational data to consider the likelihood that this association could have come about simply because of a common genetic cause of both smoking and lung cancer; they developed a sensitivity-analysis technique to show that confounding by a genetic factor was unlikely to completely account for the association. More generally, if, in an era of personal genetics, information is available on the extent to which genetic self knowledge affects behavioral and environmental exposures and the extent to which specific genetic factors increase the risk of specific diseases, it may then be possible to use sensitivity-analysis techniques[14,29,31–34] to assess the likelihood that associations arising from observational data might be due to or altered by genetic confounding.

A simple rule of thumb can be useful in this regard. Suppose we are interested in the association between an exposure and an outcome and have not adjusted for a dichotomous covariate U indicating the presence or absence of risk-elevating alleles. Let γ denote the risk ratio for the outcome comparing U = 1 and U = 0 conditional on the exposure and measured covariates, and let π_1 and π_0 denote the prevalence of U among the exposed and unexposed subjects, respectively. The ratio between the estimate for the effect of the exposure on the outcome obtained from the data and that which would have been obtained had adjustment been made for the unmeasured variable U is given by:[32–34]

$$\frac{1+(\gamma-1)\pi_1}{1+(\gamma-1)\pi_0}$$

Thus, in the MI example in the previous section, if we had $\gamma = 1.6$, $\pi_1 = 0.2$, and $\pi_0 = 0.8$ (so that $\pi_1 - \pi_0 = -0.6$), the formula above would give 0.76, indicating that an actual risk ratio of 1.3 would be reduced by genetic confounding to $1.3 \times 0.76 \approx 1$. The formula above also holds for odds ratio when the outcome is rare. It gives only a simple rule of thumb and holds only under simplifying assumptions. Other, more sophisticated sensitivity-analysis techniques are also available in the literature.[32–34]

Another possible approach to address confounding in epidemiologic research in an era of genetic self knowledge is to restrict studies to populations in which genetic knowledge is not being utilized. Advances in genetic research and access to individual genomic information are likely to propagate to different segments of the population at different rates. Communities that, for social or religious reasons, are committed to not making use of such information may

be particularly valuable in conducting epidemiologic research without requiring the collection of genomic information, though questions of generalizability will then also be important to consider. Even aside from isolated populations, genetic testing raises issues of generalizability: if individuals who use genetic tests are also more likely to participate in studies, generalizability of study findings will be partially compromised. Another strategy to circumvent possible genetic confounding would be to continue to use existing and maturing cohorts and studies conducted before genetic self knowledge is widespread. Such prior studies may prove valuable in the future but will also raise questions of generalizability as time passes and as sociodemographic distributions change.

It should finally be noted that although genetic knowledge may alter what is required in the conduct of observational research in epidemiology, such genetic knowledge will generally not raise similar issues in the study of drug efficacy in which the study design itself randomizes the administration of treatment so that it is, at least in large samples on average, independent of genetic factors and personal genetic knowledge. Indeed, such randomized trials may be able to utilize our expanding genetic knowledge so as to effectively evaluate specific treatments or regimes tailored to individual genetic information.[5,35]

Concluding Remarks

As we have seen, the possibility that individuals will use genetic self knowledge to make decisions about behavioral and environmental exposures may introduce new genetic sources of confounding in studies of environmental exposures where none was previously present. Studies of environmental factors that do not control for genetic self knowledge may produce effect estimates biased downwards. Future association studies of genetic variants may find considerably attenuated effects after personal genetic knowledge becomes more widely available. Ascertainment bias may be present when individuals self report phenotype. Likewise, validity issues may arise with regard to study designs and analytic techniques that require a gene-environment independence assumption that may be violated by individuals' use of genetic knowledge to make decisions about environmental exposures. As genetics research advances, this phenomenon may become increasingly widespread. It is thus not only the case that our genetic knowledge will potentially revolutionize personalized medicine but that personalized genetic medicine may itself significantly alter what is necessary in the practice of research, and it will be important to be prepared for the changes that may come.

Acknowledgments

This work was supported by NIH grant R01 ES017876. The author thanks the editors and two anonymous referees for helpful comments.

References

1. Davey Smith, G., Ebrahim, S., Lewis, S., Hansell, A.L., Palmer, L.J., and Burton, P.R. (2005). Genetic epidemiology and public health: hope, hype, and future prospects. Lancet *366*, 1484–1498.
2. Khoury, M.J., Gwinn, M., Yoon, P.W., Dowling, N., Moore, C.A., and Bradley, L. (2007). The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genet. Med. *9*, 665–674.
3. McGuire, A.L., and Burke, W. (2008). An unwelcome side effect of direct-to-consumer personal genome testing: raiding the medical commons. JAMA *300*, 2669–2671.
4. Katsanis, S.H., Javitt, G., and Hudson, K. (2008). Public health. A case study of personalized medicine. Science *320*, 53–54.
5. Flockhart, D.A., Skaar, T., Berlin, D.S., Klein, T.E., and Nguyen, A.T. (2009). Clinically available pharmacogenomics tests. Clin. Pharmacol. Ther. *86*, 109–113.
6. Guttmacher, A.E., McGuire, A.L., Ponder, B., and Stefánsson, K. (2010). Personalized genomic information: preparing for the future of genetic medicine. Nat. Rev. Genet. *11*, 161–165.
7. Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature *456*, 18–21.
8. Hirschhorn, J.N. (2009). Genomewide association studies—illuminating biologic pathways. N. Engl. J. Med. *360*, 1699–1701.
9. Kraft, P., and Hunter, D.J. (2009). Genetic risk prediction—are we there yet? N. Engl. J. Med. *360*, 1701–1703.
10. Kraft, P., Wacholder, S., Cornelis, M.C., Hu, F.B., Hayes, R.B., Thomas, G., Hoover, R., Hunter, D.J., and Chanock, S. (2009). Beyond odds ratios—communicating disease risk based on genetic profiles. Nat. Rev. Genet. *10*, 264–269.
11. Check Hayden, E. (2008). How to get the most from a gene test. Nature *456*, 11.
12. Couzin, J. (2008). Genetics. DNA test for breast cancer risk draws criticism. Science *322*, 357.
13. Eriksson, N., Macpherson, J.M., Tung, J.Y., Hon, L.S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I., and Mountain, J. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet. *6*, e1000993.
14. Arah, O.A., Chiba, Y., and Greenland, S. (2008). Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. Ann. Epidemiol. *18*, 637–646.
15. Khoury, M.J., and Flanders, W.D. (1996). Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am. J. Epidemiol. *144*, 207–213.
16. Yang, Q., Khoury, M.J., and Flanders, W.D. (1997). Sample size requirements in case-only designs to detect gene-environment interaction. Am. J. Epidemiol. *146*, 713–720.
17. Albert, P.S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions. Am. J. Epidemiol. *154*, 687–693.
18. Schmidt, S., and Schaid, D.J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. Am. J. Epidemiol. *150*, 878–885.
19. VanderWeele, T.J., Hernández-Díaz, S., and Hernán, M.A. (2010). Case-only gene-environment interaction studies: when does association imply mechanistic interaction? Genet. Epidemiol. *34*, 327–334.

20. Umbach, D.M., and Weinberg, C.R. (2000). The use of case-parent triads to study joint effects of genotype and exposure. Am. J. Hum. Genet. *66*, 251–261.
21. Chatterjee, N., and Carroll, R.J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika *92*, 399–418.
22. Mukherjee, B., Zhang, L., Ghosh, M., and Sinha, S. (2007). Semiparametric Bayesian analysis of case-control data under conditional gene-environment independence. Biometrics *63*, 834–844.
23. Vansteelandt, S., VanderWeele, T.J., Tchetgen, E.J., and Robins, J.M. (2008). Multiply robust inference for statistical interactions. J. Am. Stat. Assoc. *103*, 1693–1704.
24. Hung, R.J., McKay, J.D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., et al. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature *452*, 633–637.
25. Thorgeirsson, T.E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K.P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature *452*, 638–642.
26. Amos, C.I., Wu, X., Broderick, P., Gorlov, I.P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., et al. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat. Genet. *40*, 616–622.
27. Chanock, S.J., and Hunter, D.J. (2008). Genomics: when the smoke clears…. Nature *452*, 537–538.
28. Wacholder, S., Chatterjee, N., and Caporaso, N. (2008). Intermediacy and gene-environment interaction: the example of CHRNA5-A3 region, smoking, nicotine dependence, and lung cancer. J. Natl. Cancer Inst. *100*, 1488–1491.
29. Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B., and Wynder, E.L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. J. Natl. Cancer Inst. *22*, 173–203.
30. Fisher, R.A. (1958). Lung cancer and cigarettes. Nature *182*, 108–109.
31. Flanders, W.D., and Khoury, M.J. (1990). Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. Epidemiology *1*, 239–246.
32. Rothman, K.J., Greenland, S., and Lash, T.L. (2008). Modern Epidemiology, Third Edition (Philadelphia: Lippincott Williams and Wilkins).
33. Lash, T.L., Fox, M.P., and Fink, A.K. (2009). Applying Quantitative Bias Analysis to Epidemiologic Data (New York: Springer).
34. VanderWeele, T.J., and Arah, O.A. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments and confounders. Epidemiology, in press.
35. Daly, A.K. (2010). Genome-wide association studies in pharmacogenomics. Nat. Rev. Genet. *11*, 241–246.