A self-fulfilling prophecy: are we underestimating the role of the environment in gene–environment interaction research?

Paolo Vineis

Serious mistakes have been made in the past by underestimating the effects of environment and overestimating the effects of genes.1–3 A first seminal 1972 paper by Lewontin1 drew the attention of researchers on the mistakes of partitioning nature and nurture. More recently, Kittles and Weiss, considering the definition of ‘race’, showed the lack of an obvious correspondence between genotypes and phenotypes.3

Many investigations on gene–environment interactions (GEI) are under way in different parts of the world, a subject that also appears as one of the leading items in grant calls from the National Institutes of Health (NIH) or the European Union (EU). Some on-going studies are extremely large (e.g. European Prospective Study into Cancer and Nutrition [EPIC], UK Biobank). All of them employ similar methods for genotyping, while exposure assessment is extremely variable, being for example state-of-the-art for dietary intake in EPIC, but not in other studies or for other exposures. GEI imply studying both environmental exposures (e.g. to pesticides or environmental tobacco smoke) and genetic variants that are supposed to modulate the effects of the former. However, there is an asymmetry between the two. Genotyping is in fact much more accurate than the vast majority of methods used to measure environmental exposures. This implies a lower degree of classification error, that in turn means an easier identification of associations with disease. A further difficulty is related to the rarity of many environmental exposures (that, however, may have an important impact on human health), while several of the polymorphic alleles that are investigated are extremely common (e.g. 40–50% for NAT2 or GSTM1). This, again, increases the probability of detecting an association with genotypes (if this is real), but not with environmental exposures.

Let us consider the example in Table 1. The Table refers to the exposure/genotype by different assessors and a reference standard, and the resulting observed relative risks under different conditions of classification error. For example, a classification error of 10% implies the drop of a relative risk of 2.5 to 2.3, i.e. little change. With a classification error of 90% (assessor 1), however, even a relative risk of 2.5 becomes 1.1, i.e. undetectable with common epidemiological methods. Unfortunately, while in genotyping we are more frequently in the situation of assessor 4, implying a small underestimation of risks, in the field of environmental exposures we are more frequently in the situation of 3 or even 2.

Things become even more complex if we want to study interaction, for example between a frequent exposure (prevalence 25%) and a frequent genotype (prevalence 50%). Let us suppose that classification error is 20% for the environmental exposure (sensitivity = 80%), a value very likely to be smaller than in reality for most exposures. Classification error could to be around 7% for genotyping (sensitivity 93%). This is realistic, since genotyping techniques are currently validated and extremely accurate. The consequence of this situation is that we would need approximately 1800 cases to observe main effects (but no statistical interaction between exposure and genes), if no classification error occurs; 2700 if exposure is incorrectly classified 20% of the times; and 3200 if also the genotype is mistaken 7% of the times. We consider sensitivity in the example; with specificity lower than 100% numbers increase further. They also increase further under the assumption of a statistical interaction (i.e. departure from a multiplicative model) between the gene and the environment. The issue of how sample size changes for gene–environment interactions, depending on the model of statistical interaction we choose, is discussed by Clayton and McKeigue.4

According to estimates, the common genotyping method Taqman has 96% sensitivity and 98% specificity, thus allowing little error in classification. On the contrary, sensitivity in environmental exposure assessment is quite often lower than 70% and specificity even lower. This situation is not due to a

Table 1 Effects of random classification error on relative risk estimates. R = correlation coefficients between the measurement of exposure/genotype by different assessors and a reference standard, and the resulting observed relative risks (modified from Hankinson et al.)

<table>
<thead>
<tr>
<th>Assessor</th>
<th>R</th>
<th>True relative risks (RRt)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>1</td>
<td>0.10</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>0.60</td>
<td>1.3</td>
</tr>
<tr>
<td>3</td>
<td>0.80</td>
<td>1.4</td>
</tr>
<tr>
<td>4</td>
<td>0.90</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Observed RR = exp (ln RR t * R)
poor state-of-the-art of environmental science, but to objective problems in reconstructing exposures in free-living populations with great variability and changes in time. Exposure assessment is usually difficult, involves recall of complex information—such as diet—or extrapolation from few points in space and time—such as air pollution data. If we aim at understanding the impact of rare exposures (but important for those exposed) then difficulties increase further.

The situation in genetics is still more complicated because of two further reasons. First, multiple genetic testing is becoming the norm. Thus, with the usual P-value thresholds (0.05 or 0.01) a large number of false positives is going to be found. Alternatives to common P-value thresholds, to be used in multiple genetic testing, have been discussed by Colhoun et al.5 Second, one point of view that is held by some authors (e.g. ref. 4) is that case-control studies or even case-only studies are perfectly fine to investigate genetic susceptibility. This is correctly based on the assumption that the genotype does not change in the course of time and with the onset of disease, and is not affected by recall bias. However, the weakest aspect of such study designs, at least for diseases with long latency, is again exposure assessment.

The calculations above are common sense in epidemiology, but seem to be of little concern for those who plan large studies on gene–environment interactions. It can be predicted that such studies will come up with a number of genetic associations, not so much affected by classification error, and very few credible environmental associations. This is further complicated by the fact that the vast majority of genetic polymorphisms are believed to act through interaction with exposures, so that it will be difficult also to make sense of genetic observations, if the environmental component is weak. What is the solution? The only solution I foresee is to empower exposure assessment, by investing (much more than many investigators have done until now) in strong and validated exposure assessment procedures. This means that epidemiologists should collaborate not only with geneticists, but also with environmental scientists. Large efforts have been made in the field of nutrition, but not yet in all other areas of human exposures. Goals for activities aiming at improving exposure assessment include: repeated measures, allowing assessment of regression dilution bias; and validation of novel research methods, for example metabolomics or the identification of specific DNA adducts, to detect signatures left within body fluids by metabolic processes and/or external exposures. On the other side, the study of genetic influences can be used to shed light on relevant environmental exposures, and the magnitude of their effects in population subgroups. This is the ‘Mendelian randomization’ paradigm, i.e. the idea that:

the association between a disease and a polymorphism that mimics the biological link between a proposed exposure and disease is not generally susceptible to the reverse causation or confounding that may distort interpretations of conventional observational studies ... Mendelian randomization—the random assortment of genes from parents to offspring that occurs during gamete formation and conception—provides one method for assessing the causal nature of some environmental exposures.6

This would be a different way of looking at the contribution of genes, although a limitation of this approach is that one needs to know much about the functional meaning of genetic variants to use them for the interpretation of environmental exposures. So, Mendelian randomization is not exactly an antidote to the concerns I have expressed.

Acknowledgement
I am grateful to George Davey Smith for useful suggestions.

References