There are many theories on the aetiology of multiple sclerosis but most lack much in the way of direct experimental support. Despite overwhelming evidence that the pathogenesis involves an interplay between genetic and environmental effects, progress in identifying the relevant risk factors has been painfully slow and there is practically no understanding about how these might interact and result in disease. In this issue, Hedström et al. (2011) present an integration of a specific genetic and environmental effect. In a case (n=843) control (n=1209) study carefully designed to reduce potential confounding, they consider established genetic factors (DRB1 and HLA-A) in the context of smoking; and find evidence that the effects on susceptibility attributable to these genetic loci are amplified by this lifestyle exposure. On the basis of this observation, and by analogy with established pathogenic mechanisms in rheumatoid arthritis (Karlson et al., 2010), they argue that the post-translational modification of lung proteins caused by smoking might lead to multiple sclerosis in genetically susceptible individuals by priming circulating lymphocytes against cross-reacting CNS antigens. This is an intriguing hypothesis, which if confirmed would represent a major step forward in our understanding.

Attempting to unravel the pathogenesis of any complex trait through association analysis is a perilous exercise, littered with traps for the unwary. Past experience has shown that researchers using this approach need to pay obessional attention to detail if they are to avoid adding to the enormous literature confidently describing what turn out to be false positive associations (Ioannidis, 2005). At face value nothing could be simpler than establishing the relevance of an exposure by identifying a difference in its frequency between cases and controls. In this deceptively straightforward paradigm, establishing statistical significance appears to be the most challenging issue; and simultaneously avoiding confounding due to systematic but irrelevant differences between cases and controls ought not to be difficult. Unfortunately, ensuring that the tested cases and controls have been ‘randomly’ selected from the ‘same population’ and are therefore ‘well matched’ is invariably far more difficult than it seems. In reality a host of subtle effects can undermine this process and thereby generate an apparent difference in the frequency of a tested exposure that does not relate to the pathogenesis. In this context, it is important to remember that statistical analysis only allows us to judge whether an observed difference exceeds that expected by chance. It does not of itself show that the tested exposure is relevant to the pathogenesis, nor that it is correlated with something of relevance to it, unless we can be sure that other potential causes for the observed difference have been avoided. No matter how extreme the statistical significance, confidence that any observed difference in exposure is relevant depends upon the experimental design being sufficiently robust to have excluded any potential confounding.

The efforts required to achieve an adequate study depend upon the nature of the exposure tested. Any tendency for an exposure to influence which individuals are included in the study (selection bias), or for the outcome of interest to alter the tested exposure (reverse causality), have the potential to generate false positive results. The ease with which the influence of these potential confounders can be prevented represents a fundamental, but perhaps under-recognized, advantage of association studies considering genetic rather than environmental factors. Because genotype is relatively easy to determine, unlikely to influence selection and fixed over time, studies considering genetic factors are generally highly robust and relatively easy to perform. In contrast, researchers considering environmental factors are invariably faced with trying to measure something that varies over time, while simultaneously ensuring that the exposure has had no effect on recruitment is as complete as possible are logical ways to avoid the impact of these potential confounders but surprisingly difficult to achieve. Even after making Herculean efforts, Hedström et al. (2011) were only able to obtain smoking data in 70% of their cases and 40% of controls.

In principle, replication should be a robust way to confirm the validity of an association. If different investigators have tested a particular exposure in independent samples and come up with the same result, true association would seem to be the logical conclusion. Unfortunately, for many exposures, this logic is flawed because the subtle biases capable of generating false positive associations are inherent to the exposure and therefore likely to exert consistent influences across studies rather than representing random effects. In other words, the apparent replication of an observation may be the result of a consistent bias rather than a
true association. It is the bias that has been replicated, not the effect. There have been many studies looking at smoking in multiple sclerosis and meta-analysis has suggested an encouraging degree of consistency (Hawkes, 2007). However, the inherent challenges of measuring this exposure, while limiting the effects of confounders, make it difficult to judge whether the apparent consistency is due to a subtle uncompensated bias or a true association.

In the complex situation where several independent exposures influence an outcome, regression analysis is a useful tool since it enables their relationship to be summarized in terms of a mathematical model. However, in the absence of information about pathogenesis, and therefore knowledge of the ‘true’ underlying relationships, the choice of model employed in such an analysis is primarily guided by mathematical considerations and is essentially arbitrary from the biological perspective. For convenience it is usual to employ simple mathematical models where each exposure contributes independently and the parameters estimated from the data (the regression coefficients) have some meaningful interpretation in terms of the influence of each exposure on the outcome (see Box). Given the limited degrees of freedom in such restricted

**BOX**

Two models are commonly considered in regression analysis; the linear (linear-additive) and the logistic (linear-multiplicative). In the situation where there are \( n \) risk factors influencing an outcome the linear regression model may be written as:

\[
R = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n
\]

where \( R \) is the risk of the outcome in an individual, \( \beta_0 \) is the baseline risk (the risk in those individuals not having exposure to any of the risk factors), \( \beta_i \) is the regression coefficient for the \( i \)-th exposure, and \( x_i \) is the indicator variable for that exposure (for example, 0 if the exposure is absent and 1 if it is present, or for a biallelic genetic marker 0, 1 or 2 depending upon the number of risk alleles carried). In this model, the regression coefficients (the \( \beta \)'s) equate to the average increase in the absolute risk of the outcome for each unit increase in the respective exposure.

In contrast, the logistic model is parameterized in terms of the log of the odds of the outcome \( (O) \) rather than risk and may be written as:

\[
\ln(O) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n
\]

where \( \alpha_0 \) is log of the baseline odds (the odds of the outcome in those individuals not exposed to any of the risk factors), \( \alpha_i \) is the regression coefficient for the \( i \)-th exposure and \( x_i \) is the indicator variable for that exposure in this model the regression coefficients (the \( \alpha \)'s) may be interpreted as log of the odds ratio attributable to the respective exposure. That is, \( e^{\alpha_i} \) is the odds ratio attributable to the \( i \)-th exposure. Rearranging the logistic equation in terms of risk, we have:

\[
R = \frac{e^{z}}{e^{z} + 1}
\]

with \( z = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \)

In considering the results from regression analysis employing these simple models, it is essential to remember that both are little more than mathematical conveniences and that neither has any inherent biological meaning (Fig. 1). Although, the regression coefficients generated have a logical interpretation within the context of the particular model employed, in generating them the actual values calculated are unlikely to be more than a reflection of the average ‘real’ effect of an exposure. These calculated values should not be endowed with more meaning than they deserve. It is also worth noting that in a well-balanced study whenever an outcome is determined by the influence of multiple modest effects, there is unlikely to be much difference in the extent of fit between a linear or a logistic model.

The binary nature of disease status (individuals either have or do not have multiple sclerosis) means that in the analysis of case–control studies a logistic model will almost always provide a better mathematical fit than a linear model.

---

**Figure 1** The left hand panel shows a weight suspended by a chain, where each link in the chain has some probability of failing—\( p_i \). In this situation the linear model, the arithmetic sum of these individual risks, will provide a reasonable estimate for the risk of the weight failing (the outcome) as long as the individual risks are small enough to allow us to ignore the possibility that two or more links might fail simultaneously. In this situation, the linear model has a meaningful interpretation since it approximates to the outcome being determined by a set of independent sufficient causes (i.e. it equates to heterogeneity of causation). Any one of many possible causes (exposures) may result in the outcome. Conversely the right hand panel shows a weight suspended by a rope, where each strand of the rope has a probability of failure—\( p_i \). In this situation, the probability of the weight failing is given by the product of the individual probabilities \( (p_i) \) rather than their sum. If the probability of the outcome is low, for each strand (exposure) \( p_i \) will be \( \leq 1 - p_i \) and therefore there will be little difference between odds and risk (and hence between odds ratio and relative risk). Under the rare disease hypothesis the logistic model thus approximates to this multiplicative model. Although the multiplicative model does not have a simple mechanistic interpretation akin to that seen with the linear model, analogy with the figure suggests that it might equate to causation which requires simultaneous failure in a number of systems, as expected if the disease resulted from failure in a biological process which is highly redundant.
models, we can anticipate that in many situations an accurate fit to the data will only be possible when these simple models are expanded to include terms based on the product of various exposures. Within the context of the model these ‘interaction’ terms may be interpreted as the quantitative change in risk ascribed to one exposure when that occurs in the presence of another. However, in the absence of any biological correlate for the model being employed, these statistical interaction terms do not necessarily indicate the presence of biologically meaningful (qualitative) interactions (that is, epistasis). The statistical rather than biological nature of these terms can be appreciated by considering the situation in which a simple logistic model provides an accurate fit to a set of data. In this case it is likely that interaction terms will need to be included if a linear model is to be equally well fitted to the same data, and vice versa (Clayton, 2009). A biological meaning is improbable for terms that may disappear, or even reverse, simply by changing the model chosen to form the foundation of the analysis (Clayton, 2009). In short, the demonstration of a significant interaction term in a regression analysis based on a convenient mathematical model is evidence that the tested data are inconsistent with the model employed; but is not necessarily evidence for a biologically meaningful interaction regardless of whether the tested model is linear or logistic. Hedström et al. (2011) base their analysis on a linear model which amounts to testing whether DRB1*15, HLA-A*X (i.e. not A*02) and smoking are independent sufficient causes for multiple sclerosis. While this may be a biologically tangible model, it is debatable whether it is a reasonable or logical starting position.

In order to reduce the degrees of freedom in their main analysis and thereby maximizing the power to detect interaction, Hedström et al. (2011) consider each of the genetic loci studied (DRB1 and HLA-A) in terms of allele carriage rather than individual genotypes. While this is a legitimate strategy, it is disappointing that statistically significant interaction is only apparent when the HLA-A heterozygotes (02/X) are grouped with the low-risk homozygotes (02/02) and not when these heterozygotes are grouped with the high-risk heterozygotes (X/X).

In summary, it is very hard to perform reliable association studies assessing genetic factors, but it is even harder to perform such studies for environmental exposures, where vulnerability to confounding is increased and measurement is inherently less precise. Over the next few years, we can hope to see the analysis of environmental effects expand and anticipate that claims for significant interaction (gene–gene and gene–environment) will follow. If we are to avoid another round of false positive and misleading results these studies need to be at least as rigorous as those we are now used to seeing in genetic analysis. Great care must be taken to distinguish statistical from biological interaction. The work of Hedström et al. (2011) illustrates the level of effort and attention to detail required in such studies and probably represents the best that can be achieved in the context of a case–control analysis. Although, their reported interaction is critically dependent upon how heterozygotes are handled and only carries a modest bottom line P-value (2%), it seems illogical to conclude that their proposed pathogenic hypothesis stands or falls on the basis of whether or not there is interaction. Surely the potential for the proposed mechanism primarily depends upon the reliability of the evidence that smoking really is a risk factor in multiple sclerosis. In other words, are the enormous and detailed efforts made by these authors to control for confounding adequate? Without doing a cohort study where smokers and non-smokers are followed over time and the incidence of multiple sclerosis is determined prospectively, it is hard to see how the study could be substantially improved. However, since the extent to which they have succeeded in controlling bias cannot be directly validated, it is ultimately up to each reader to decide whether or not he or she is convinced that the effects of potential confounders such as selection bias and reverse causality have been adequately excluded. But regardless of interpretation, the medical advice is clear: ‘don’t smoke, it’s bad for you’.

Funding
This work was supported by the Wellcome Trust (084702) and the Cambridge NIHR Biomedical Research Centre.

Stephen Sawyer1 and Garrett Hellenthal2
1Department of Clinical Neurosciences, University of Cambridge, Addenbrooke’s Hospital, Cambridge, CB2 0QQ, UK
2The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

Advance Access publication February 9, 2011
doi:10.1093/brain/awq384

References


