Interaction Between Genetic and Environmental Factors in Multiple Sclerosis?

Paul A. Torvund *

September 18, 2015

Abstract

Genetic, as well as environmental factors, may represent risk factors for multiple sclerosis. A question of particular interest is whether risk factors may work synergistically. The main findings in this study are that whereas two genetic risk factors and a history of smoking have highly significant main effects on the risk (P-value <0.001), there is not sufficient evidence to proclaim an interaction between any of these genetic factors and smoking at a 5% level of significance. However, there is a significant interaction between gender and smoking. The interaction indicates that smoking among males contributes more to the risk of getting multiple sclerosis than smoking among females. Previous Epstein Barr virus infection is a significant risk factor when adjusting for smoking only. The data reveal a strong protective effect of snuffing. Contrary to the estimate of the effect of Epstein Barr virus infection, the latter estimate is statistically significant also after inclusion of several adjustment variables.

Contents

*This article is a medical student thesis submitted at the Faculty of Medicine, Univeristy of Oslo. Thanks to Professor Petter Laake for supervision and Professor Hanne Harbo and Research Fellow Christian Page for providing the data and commenting on a previous version.

1 Introduction

Multiple sclerosis is the most common autoimmune disorder that affects the central nervous system. It is classified as a demyelinating condition. Demyelination causes diminished or extinguished propagation of signals along the axons of neurons. The condition can lead to a variety of physical and sometimes cognitive symptoms. There is no cure. Treatment is symptomatic and focused on prevention of inflammatory attacks and delay of progression (Linker et al. [7]).

Multiple sclerosis is considered to be an autoimmune disease, although no autoantigen has been identified (McFarland and Martin [8]). The causes of the autoimmunity are not determined. Genetic, as well as environmental factors may contribute to the susceptibility to multiple sclerosis. A relatively strong genetic risk factor seems to be the presence of the Human Leukocyte Antigen DRB1*15 allele. Absence of the Human Leukocyte Antigen A*02 allele is thought to provide protection (Lincoln et al. [6] and Brynedal et al. [2]). Among the candidates for environmental risk factors are a history of smoking, passive smoking or snuffing, low serum levels of vitamin D, and infectious disease caused by the Epstein Barr virus. It is also well known that women are more at risk of getting multiple sclerosis than men (Alonso and Hernan [1]).

A hypothesis of particular interest to this study is that environmental factors such as smoking may prime the immune system to degrade myelin or cells that produce myelin. If that is the case, one would suspect statistical interaction between enironmental and genetic factors. It should be noted that such interaction could exist for other reasons than the hypothesized priming mechanism.

To some extent the present study replicates a study conducted by Hedström et al. published in 2011 in *Brain* [3]. Both are case control studies with genetic and environmental data gathered for a group of cases consisting of people diagnosed with multiple sclerosis, and a control group extracted from the general population. In particular, Hedström et al. analyze, using logistic regression, the effects on the probability of getting multiple sclerosis of the following variables: a history of smoking, the presence of the Human Leukocyte Antigen DRB1*15 allele and the lack of the Human Leukocyte Antigen A*02 allele. In addition, they consider all interactions there can be between these variables. They also state that they adjust for age, gender, residential area and ancestry. There are no data for ancestry and residential area in the prestent study.

Hedström et al. consider several additional genetic factors, but their study found presence of the Human Leukocyte Antigen DRB1*15 allele to

be the most important genetic risk factor, and presence of the Human Leukocyte Antigen A*02 allele to be the most important genetic protective factor. The genetic factors considered in the present study are restricted to these two genetic factors. Both data sets include binary smoking data.

2 Material and Method

2.1 Material

The project is designed as a case control study. The sample of multiple sclerosis cases consists of 530 individuals, collected and genotyped by the Multiple Sclerosis Research Group at Oslo University Hospital (ous-research.no/harbo). The sample of controls, which consists of 918 indivuals, is recruited through the Norwegian Bone Marrow Registry in collaboration with Professor Benedicte A. Lie. In total, 1448 individuals are included in the study. In comparison, the Hedström et al. study [3] contains 843 multiple sclerosis cases and 1209 controls.

Generally, for each person i in the data set, the following is registered:

$$MS_i = \begin{cases} 0 & \text{if individual } i \text{ does not have multiple sclerosis,} \\ 1 & \text{if individual } i \text{ has multiple sclerosis.} \end{cases}$$

This is the dependent variable. We shall consider seven explanatory variables. These are:

$$HLA\_DRB1\_15_i = \begin{cases} 0 & \text{if individual } i \text{ is not carrier of} \\ & \text{Human Leukocyte Antigen DRB1*15,} \\ 1 & \text{if individual } i \text{ is carrier of} \\ & \text{Human Leukocyte Antigen DRB1*15.} \end{cases}$$

Being a carrier is thought to be a genetic risk factor. This factor is registered for 1227 of the 1448 indivuduals.

$$HLA\_A\_02_i = \begin{cases} 0 & \text{if individual } i \text{ is carrier of} \\ & \text{Human Leukocyte Antigen A*02,} \\ 1 & \text{if individual } i \text{ is not carrier of} \\ & \text{Human Leukocyte Antigen A*02.} \end{cases}$$

Being a carrier is thought to be a genetic protective factor. This factor is registered for 1146 of the 1448 indivuduals.

$$Gender_i = \begin{cases} 0 & \text{if individual } i \text{ is male,} \\ 1 & \text{if individual } i \text{ is female.} \end{cases}$$

This item is registered for all individuals that are included in the study.

$$Epstein\_Barr_i = \begin{cases} 0 & \text{if individual } i \text{ has multiple sclerosis and has not had an Epstein Barr virus infection prior to getting multiple sclerosis, or if individual } i \text{ does not have multiple sclerosis and has never had an Epstein Barr virus infection,} \\ 1 & \text{otherwise.} \end{cases}$$

This factor is registered for 1306 of the 1448 indivuduals.

$$Smoker_i = \begin{cases} 0 & \text{if individual } i \text{ has multiple sclerosis and has never smoked prior to getting multiple sclerosis, or if individual } i \text{ does not have multiple sclerosis and has never smoked,} \\ 1 & \text{otherwise.} \end{cases}$$

This factor is registered for 1422 of the 1448 indivuduals.

$$Passive\_Smoker_i = \begin{cases} 0 & \text{if individual } i \text{ has multiple sclerosis and has never been a passive smoker prior to getting multiple sclerosis, or if individual } i \text{ does not have multiple sclerosis and has never been a passive smoker,} \\ 1 & \text{otherwise.} \end{cases}$$

This factor is registered for 1431 of the 1448 indivuduals.

$$Snuffer_i = \begin{cases} 0 & \text{if individual } i \text{ has multiple sclerosis and has never been a snuffer prior to getting multiple sclerosis, or if individual } i \text{ does not have multiple sclerosis and has never been a snuffer,} \\ 1 & \text{otherwise.} \end{cases}$$

This factor is registered for 1432 of the 1448 indivuduals.

In addition, the data set contains information about person i's age, person i's age at onset of the disease, person i's Expanded Disability Status Score, as well as whether the disease is of the Relapsing Remitting/Secondary Progressive type or the Primary Progressive type. These information items will not be used, due to reasons that will be considered in section 4, *Discussion.*

2.2 Method

The data are analyzed using regression analysis, which provides a way to estimate the effects of the explanatory variables on the dependent variable. Regressions can be used to calculate a prognosis for the dependent variable given the values that the explanatory variables may take. However, that an explanatory variable influences the dependent variable cannot in general be interpreted as a causal effect. One main reason is that the explanatory variable in general will be correlated with other variables that are at least in part responsible for the causality. If these other variables are not included as explanatory variables in the model, the effect estimate of the explanatory variable under consideration will be biased. This bias is called omitted variable bias. Including the variables that (partly) are responsible for the causality, is called adjusting (controlling) for these variables. If one is able to do that properly, regression analysis is a method that in principle can be used to assess causal effects based on observational data. It should be kept in mind, however, that it is generally unknown what controls one should use. Furthermore, there are other sources of bias, the most important of which may be sampling bias, measurement error, and reverse causality.

MS is the dependent variable in the regression analyses that will be performed. Possible explanatory variables are $Gender$, $Smoker$, HLA_A_02, HLA_DRB1_15, $Passive\_Smoker$, $Epstein\_Barr$, and $Snuffer$, as well as product terms. Recall that the presence in a person's genome of the Human Leukocyte Antigen A*02 allele is considered to be a protective factor, so that not having that is considered to be a risk factor. Therefore, all the variables take the value 1 only if they if they are thought *a priori* to contribute to the risk of multiple sclerosis.

Let Y be the dependent variable, and $X_1, X_2, \ldots, X_k$ explanatory variables. The purpose of regression analysis is to approximate the unknown regression function $E(Y \mid X_1, X_2, \ldots, X_k)$. The justification for using regression analysis on a data set with binary dependent variable is that if Y takes the values 0 and 1 only, then its expected value is the probability that $Y = 1$, i.e.,

$$E(Y \mid X_1, X_2, \ldots, X_k) = Pr(Y = 1 \mid X_1, X_2, \ldots, X_k). \tag{1}$$

If one uses linear regression, one postulates that the best model for $E(Y \mid X_1, X_2, \ldots, X_k)$ is

$$E(Y \mid X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k. \tag{2}$$

The parametres $\beta_0, \beta_1, \ldots, \beta_k$ are estimated using the data. One advantage of this approach is that the interpretation of $\beta_i, i \in \{0, \ldots, k\}$ is particularly simple. If $i \neq 0$, β_i equals the increase in the probability of the dependent variable Y as a result in a unit's increase in the explanatory variable X_i.

The major drawback of linear regression when the dependent variable is binary is that linear regression may not fit the data well, especially outside the range where there is a high density of observations. It may not be clear, however, that that is a reasonable objection in cases where the explanatory variables are also binary, as is the case here. A more serious objection may be that estimates obtained by linear regression can be biased due to sample selection, which is an obvious concern in case control studies.

A major alternative to linear regression if the dependent variable is binary is logistic regression (see Hosmer and Lemeshow [5] for an introduction). If one uses binary regression, one postulates that the best model for $E(Y \mid X_1, X_2, \ldots, X_k)$ is

$$E(Y \mid X_1, X_2, \ldots, X_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}} \tag{3}$$

To simplify notation, let

$$Pr(Y = 1 \mid X_1, X_2, \ldots, X_k) = P(Y) \tag{4}$$

and let $t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. Recalling equation 1 we now have that

$$ln\left(\frac{P(Y)}{1 - P(Y)}\right) = ln\left(\frac{\frac{1}{1+e^{-t}}}{1 - \frac{1}{1+e^{-t}}}\right) = ln\left(\frac{1}{1 - e^{-t} - 1}\right) = ln\left(e^{t}\right) = t. \tag{5}$$

This means that the odds $\frac{P(Y)}{1-P(Y)}$ equals $e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}$.

Consider now two states, one in which $X_i = 1$ and one in which $X_i = 0$, while all other explanatory variables remain constant. Denote by $Pr(Y \mid X_i = 1)$ the probability that $Y = 1$ given that $X_i = 1$ and by $Pr(Y \mid X_i = 0)$ the probability that $Y = 1$ given that $X_i = 0$. Now the *odds ratio*

$$OR = \frac{\frac{Pr(Y|X_i=1)}{1-Pr(Y|X_i=1)}}{\frac{Pr(Y|X_i=0)}{1-Pr(Y|X_i=0)}} \tag{6}$$

has a relatively simple interpretation: individuals having $X_i = 1$ (i.e., individuals exposed to the characteristic signified by X_i) are OR times more likely to have $Y = 1$ (e.g., to be sick) than indivuduals having $X_i = 0$ (i.e., individuals not exposed to the characteristic signified by X_i). Furthermore, the discussion above gives that OR can be expressed as

$$OR = \frac{e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_i + \ldots + \beta_k X_k)}}{e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_{i-1} X_{i-1} + \beta_{i+1} X_{i+1} + \ldots + \beta_k X_k)}} = e^{\beta_i}, \tag{7}$$

so when β_i (and thereby e^{β_i}) is estimated using logistic regression, we have a ready interpretation of that number.

This can readily be generalized to the situation where in one state, variables $X_{m_1}, \ldots, X_{m_n}$, $m_1, \ldots, m_n \in \{0, \ldots, k\}$, are all equal to 1, and in the other state, the same variables $X_{m_1}, \ldots, X_{m_n}$, are all equal to 0, while all other variables $X_j, j \notin \{m_1, \ldots, m_n\}$ remain the same for both states. Denote by $Pr(Y \mid X_{m_1} = 1, \ldots, X_{m_n} = 1)$ the probability that $Y = 1$ given that the first of these states occurs, and by $Pr(Y \mid X_{m_1} = 0, \ldots, X_{m_n} = 0)$ the probability that $Y = 1$ given that the second of these states occurs. Now the odds ratio between the first and the second state is given by

$$OR = \frac{\frac{Pr(Y|X_{m_1}=1,\ldots,X_{m_n}=1)}{1-Pr(Y|X_{m_1}=1,\ldots,X_{m_n}=1)}}{\frac{Pr(Y|X_{m_1}=0,\ldots,X_{m_n}=0)}{1-Pr(Y|X_{m_1}=0,\ldots,X_{m_n}=0)}}, \tag{8}$$

and OR can now be expressed as

$$OR = e^{(\beta_{m_1}+\ldots+\beta_{m_n})}. \tag{9}$$

We will now comment on the problem of selection bias in case control studies. Using logistic regression, the estimated coefficients $\hat{\beta}_i$ of $\beta_i, i \in \{1, \ldots, k\}$ will not be vulnerable to selection bias due to the inherent structure of case control studies (but can of course be vulnerable to other sources of bias like bias due to omitted variables, reverse causality and measurement error). The intercept β_0 cannot directly be estimated without bias using data from a case control study. Because of the study design whereby cases are chosen systematically, in our study based on their having multiple sclerosis, the estimated risk will depend on the size of the case sample relatively to the control sample. The bigger the case sample, keeping the size of the control sample fixed, the higher the estimated risk. This kind of bias can, however, be corrected. An unbiased estimator $\hat{\beta}_0^*$ for β_0 is given by

$$\hat{\beta}_0^* = \hat{\beta}_0 + ln\left(\frac{\pi}{1-\pi}\right) - ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right), \tag{10}$$

where $\hat{\beta}_0$ is the estimate for β_0 given by the regression analysis, π is the true prevalence of the dependent variable in the population, and $\hat{\pi}$ is the prevalence in the sample (see web site [9], page 16).

To reduce variance in the parameter estimates, the number of controls in a study should as a rule of thumb be at least five times as large as the case group (see web site [9], page 17).

3 Results and Discussion

3.1 Regression results: unadjusted and main effects

Consider the regression model

$$ln\left(\frac{P(MS)}{1-P(MS)}\right) = \beta_0 + \beta_1 Smoker, \tag{11}$$

where the notation $P(Y)$ is defined in equation 4. The results of the regression are shown in table 1.

	$\hat{\beta}_1$	Standard Error	P-value	$e^{\hat{\beta}_1}$
Smoker	0.882	0.121	<0.001	2.416

Table 1: The unadjusted effect of smoking on the risk of getting multiple scleroisis.

Conclusion: the null hypothesis that smoking has no effect on the risk of getting multiple sclerosis is rejected based on these data.

Consider

$$ln\left(\frac{P(MS)}{1-P(MS)}\right) = \beta_0 + \beta_1 Gender, \tag{12}$$

where the notation $P(Y)$ is defined in equation 4. The results of the regression are shown in table 2.

	$\hat{\beta}_1$	Standard Error	P-value	$e^{\hat{\beta}_1}$
Gender	0.700	0.119	<0.001	2.014

Table 2: The unadjusted effect of being a female versus being a male on the risk of getting multiple scleroisis.

Conclusion: the null hypothesis that gender does not affect the risk of getting multiple sclerosis is rejected based on these data.

Consider

$$ln\left(\frac{P(MS)}{1-P(MS)}\right) = \beta_0 + \beta_1 HLA\_DRB1\_15, \tag{13}$$

where the notation $P(Y)$ is defined in equation 4. The results of the regression are shown in table 3:

	$\hat{\beta}_1$	Standard Error	P-value	$e^{\hat{\beta}_1}$
HLA_DRB1_15	1.232	0.135	<0.001	3.430

Table 3: The unadjusted effect of having the Human Leukocyte Antigen DRB1*15 allele on the risk of getting multiple scleroisis.

Conclusion: the null hypothesis that having the Human Leukocyte Antigen DRB1*15 allele does not affect the risk of getting multiple sclerosis is rejected based on these data.

Consider

$$ln\left(\frac{P(MS)}{1-P(MS)}\right) = \beta_0 + \beta_1 HLA\_A\_02, \tag{14}$$

	$\hat{\beta}_1$	Standard Error	P-value	$e^{\hat{\beta}_1}$
HLA_A_02	0.534	0.135	<0.001	1.705

Table 4: The unadjusted effect of not having the protective effect of the Human Leukocyte Antigen A*02 allele on the risk of getting multiple scleroisis.

where the notation $P(Y)$ is defined in equation 4. The results of the regression are shown in table 4:

Conclusion: the null hypothesis that not having the protective effect of Human Leukocyte Antigen A*02 allele does not affect the risk of getting multiple sclerosis is rejected based on these data.

It is clear that there may be an association (correlation) between passive smoking and smoking. Assuming that many passive smokers are also smokers, then if one wants to estimate the effect of passive smoking only on the risk of getting multiple sclerosis, one must adjust for smoking. The same goes for snuffing: assuming that it is common to snuff as an alternative to smoking, then if one is interested in the effect of snuffing only on the risk of getting multiple sclerosis, one must adjust for smoking. As for Epstein Barr virus infections, it may be that smokers are more likely to get them than non-smokers, so again, to estimate the effect of Epstein Barr virus infection only on the risk of getting MS, it may be wise to adjust for smoking. Therefore, to reduce bias, we shall adjust for the variable $Smoker$ as the effects of the variables $Epstein\_Barr$, $Passive\_Smoker$, and $Snuffer$ on the risk of getting multiple sclerosis are estimated.

Consider

$$ln\left(\frac{P(MS)}{1-P(MS)}\right) = \beta_0 + \beta_1 Smoker + \beta_2 Passive\_Smoker, \tag{15}$$

where the notation $P(Y)$ is defined in equation 4. The results of the regression are shown in table 5:

	$\hat{\beta}_i, i = 1, 2$	Standard Error	P-value	$e^{\hat{\beta}_i}$
$Smoker$	0.904	0.123	<0.001	2.469
$Passive\_Smoker$	-0.134	0.127	0.291	0.874

Table 5: The effect of passive smoking, while controlling for smoking.

There is a 29.1% probability of randomly obtaining the estimated result or a result more adverse to the null hypothesis of zero effect of passive smoking on the probability of getting multiple sclerosis. Conclusion: The null hypothesis is not rejected based on these data, and the variable $Passive\_Smoker$ will be omitted in regressions below.

Consider

$$ln\left(\frac{P(MS)}{1-P(MS)}\right) = \beta_0 + \beta_1 Smoker + \beta_2 Snuffer, \quad (16)$$

where the notation $P(Y)$ is defined in equation 4. The results of the regression are shown in table 6.

	$\hat{\beta}_i, i = 1, 2$	Standard Error	P-value	$e^{\hat{\beta}_i}$
$Smoker$	0.948	0.123	<0.001	2.581
$Snuffer$	-0.549	0.170	0.001	0.578

Table 6: The effect of snuffing, while controlling for smoking.

Conclusion: the null hypothesis that snuffing does not affect the risk of getting multiple sclerosis is rejected based on these data. Indeed, the data reveal a significant protective effect of snuffing.

If one finds no biological reason to believe in this result, one likely explanation may be omitted variable bias. The habit of snuffing could represent some other characteristic of the individual, e.g. something about the individual's social status, that could be the real protective factor. Results that indicate that snuffing does not increase the risk of getting multiple sclerosis have previously been found (Hedström et al. [4]).

Consider

$$ln\left(\frac{P(MS)}{1-P(MS)}\right) = \beta_0 + \beta_1 Smoker + \beta_2 Epstein\_Barr, \quad (17)$$

where the notation $P(Y)$ is defined in equation 4. The results of the regression are shown in table 7:

	$\hat{\beta}_i, i = 1, 2$	Standard Error	P-value	$e^{\hat{\beta}_i}$
$Smoker$	0.940	0.132	<0.001	2.559
$Epstein\_Barr$	0.489	0.165	0.003	1.631

Table 7: The effect of previous Epstein Barr virus infection, while controlling for smoking.

Conclusion: the null hypothesis that previous Epstein Barr virus infection does not affect the risk of getting multiple sclerosis is rejected based on these data.

So far we have found six variables to have significant effect on the risk of

getting multiple sclerosis. Consider the regression that includes all of them:

$$\begin{aligned} ln\left(\frac{P(MS)}{1-P(MS)}\right) =& \beta_0 + \beta_1 Smoker \\ &+\beta_2 Gender \\ &+\beta_3 HLA\_DRB1\_15 \\ &+\beta_4 HLA\_A\_02 \\ &+\beta_5 Snuffer \\ &+\beta_6 Epstein\_Barr \end{aligned} \tag{18}$$

where the notation $P(Y)$ is defined in equation 4. The results of the regression are shown in table 8.

	$\hat{\beta}_i, i = 1, \ldots, 6$	Standard Error	P-value	$e^{\hat{\beta}_i}$
$Smoker$	1.141	0.181	<0.001	3.129
$Gender$	0.451	0.184	0.014	1.570
HLA_DRB1_15	1.393	0.160	<0.001	4.029
HLA_A_02	0.640	0.160	<0.001	1.896
$Snuffer$	-0.817	0.287	0.004	0.442
$Epstein\_Barr$	0.246	0.231	0.287	1.279

Table 8: The effect of the main regressors, including $Epstein\_Barr$, but excluding interaction terms.

There is a 28.7% probability of randomly obtaining the estimated result or a result more adverse to the null hypothesis of zero effect of Epstein Barr virus infection on the probability of getting multiple sclerosis. Conclusion: The null hypothesis concerning $Epstein\_Barr$ is not rejected based on these data with this regression.

It is notable that while the statistical significance of Epstein Barr virus infection was lost as more variables were included into the regression, the estimated protective effect of snuffing is stronger in this expanded model, and still significant. The inclusion of more controls has, if anything, strengthened the hypothesis that snuffing is protective or that snuffing representes something that is protective. That the statistical significance of Epstein Barr virus infection was lost may be due to a combination of lack of statistical power and a high degree of association with one or more of the included variables. The previous estimated effect may have been mainly due to omitted variable bias that now is corrected for.

3.2 Assessment of interactions

In the search for possible interactions, we are now left with the five explanatory variables $Smoker$, $Gender$, HLA_DRB1_15, HLA_A_02, and $Snuffer$. The variable $Epstein\_Barr$ is exluded.

With five explanatory variables, there are ten possible first order interactions. All but one are insignificant in a model that includes the above mentioned five explanatory variables. The exception is the interaction between gender and smoking. The model is as follows:

$$\begin{aligned} ln\left(\frac{P(MS)}{1-P(MS)}\right) =& \beta_0 + \beta_1 Smoker \\ &+\beta_2 Gender \\ &+\beta_3 HLA\_DRB1\_15 \\ &+\beta_4 HLA\_A\_02 \\ &+\beta_5 Snuffer \\ &+\beta_6 Smoker \times Gender \end{aligned} \tag{19}$$

where the notation $P(Y)$ is defined in equation 4. The results of the regression are shown in table 9.

	$\hat{\beta}_i, i = 1, \ldots, 6$	Standard Error	P-value	$e^{\hat{\beta}_i}$
$Smoker$	1.645	0.338	<0.001	5.183
$Gender$	1.040	0.328	0.002	2.829
HLA_DRB1_15	1.372	0.149	<0.001	3.942
HLA_A_02	0.588	0.149	<0.001	1.800
$Snuffer$	-0.738	0.263	0.005	0.478
$Smoker \times Gender$	-0.798	0.385	0.038	0.450

Table 9: The effect of the main regressors, excluding $Epstein\_Barr$, but including interaction between $Smoker$ and $Gender$.

The estimates of the effects of HLA_DRB1_15, HLA_A_02 and $Snuffer$ are essentially as they have been in previous regressions. The individual effects of $Smoker$ and $Gender$ are estimated to be substantially higher than before. The interaction term $Smoker \times Gender$ is estimated to give a significant protective effect. The protective effect kicks in when $Smoker = 1$ and $Gender = 1$. What this means is that the added risk that is due to smoking is higher among men than among women, and among non-smokers, there is an added risk due to being a woman versus being a man. So, restricting one's considerations to the risk of getting multiple sclerosis, one may conclude that the data indicate that smoking is more dangerous to men than to women (but smoking is also a risk factor to women).

One should consider other possible explanations. In theory it could be for instance that men smoke more heavily than women. That would be something that is not reflected in the data. Other threats to internal validity should also be considered. The lack of data on social status may be of particular concern.

The estimated effect of the interaction term itself can never be affected by omitted variable bias as long as one controls for the variables that are included in the interaction. For instance, if one is interested in the effect of $Smoker \times Gender$ only, it is always sufficient to control for $Smoker$ and $Gender$ to be certain that any omitted variable bias is avoided. The reason is almost trivial. Any omitted variables one might think could be correlated with the interaction term must have this correlation via at least one of the terms that are included in the interaction. But if one controls for these terms, one automatically controls for any omitted variables.

3.3 A comparison to Hedström et al.

In this subsection we shall aim at reproducing two main results of Hedström et al. [3]. One main finding of theirs is formulated as follows: "Compared with non-smokers with neither of the genetic risk factors, the odds ratio was 13.5 (8.1 − 22.6) for smokers with both genetic risk factors." The corresponding odds ratio is, using the same notation as in equation 8,

$$OR = \frac{\frac{Pr(MS|Smoker=1,HLA\_DRB1\_15=1,HLA\_A\_02=1)}{1-Pr(MS|Smoker=1,HLA\_DRB1\_15=1,HLA\_A\_02=1)}}{\frac{Pr(MS|Smoker=0,HLA\_DRB1\_15=0,HLA\_A\_02=0)}{1-Pr(MS|Smoker=0,HLA\_DRB1\_15=0,HLA\_A\_02=0)}}. \tag{20}$$

Consider the following model:

$$\begin{aligned} ln\left(\frac{P(MS)}{1-P(MS)}\right) =& \beta_0 + \beta_1 HLA\_DRB1\_15 \\ &+\beta_2 HLA\_A\_02 \\ &+\beta_3 Smoker, \end{aligned} \tag{21}$$

where the notation $P(Y)$ is defined in equation 4. Omitting control variables may introduce omitted variable bias, but the estimate itself will be valid for prognostic purposes. Therefore, for the purposes of prognostics, the model 21 is kept as simple as possible. The results of this regression are shown in table 10.

	$\hat{\beta}_i, i = 1, \ldots, 3$	St. Err.	P-value	$e^{\hat{\beta}_i}$
HLA_DRB1_15	1.334	0.146	<0.001	3.797
HLA_A_02	0.608	0.147	<0.001	1.837
$Smoker$	1.005	0.162	<0.001	2.731

Table 10: The result of a regression designed to provide a prognosis of effect of smoking and the two genetic risk factors.

The effect estimate is

$$\widehat{OR} = e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3} = e^{\hat{\beta}_1}e^{\hat{\beta}_2}e^{\hat{\beta}_3} = 3.797 \times 1.837 \times 2.731 = 19.0, \tag{22}$$

which is within the confidence interval $(8.1 - 22.6)$ found by Hedström et al.

As for the confidence interal of this estimate, note that for stochastic variables $Y_i, i \in \{1, \ldots, n\}$,

$$Var\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} Var(Y_i) + 2\sum\sum_{i<j} Cov(Y_i, Y_j). \tag{23}$$

This means that the standard error SE of $(\beta_1 + \beta_2 + \beta_3)$ is given by

$$SE(\beta_1 + \beta_2 + \beta_3) = \sqrt{\widehat{Var}(\beta_1) + \widehat{Var}(\beta_2) + \widehat{Var}(\beta_3) + 2\widehat{Cov}(\beta_1, \beta_2) + 2\widehat{Cov}(\beta_1, \beta_3) + 2\widehat{Cov}(\beta_2, \beta_3)}. \tag{24}$$

For stochastic variables X and Y,

$$\widehat{Cov}(X, Y) = \widehat{Corr}(X, Y)SE(X)SE(Y). \tag{25}$$

The three factors on the right hand side are known, so this reduces to

$$\begin{aligned} &SE(\beta_1 + \beta_2 + \beta_3) \\ &= \sqrt{0.0213 + 0.0216 + 0.0262 + 0.0045 + 0.0011 + 0.0022} = 0.278. \end{aligned} \tag{26}$$

The 95% confidence interval for the odds ratio is therefore approximately given by

$$\begin{aligned} &\left(e^{(\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 - 1.96SE)}, e^{(\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + 1.96SE)}\right) = \left(\frac{19.05}{1.72}, 19.05 \times 1.72\right) \\ &= (11.1, 32.8)\,. \end{aligned} \tag{27}$$

Hedström et al.'s estimate 13.5 is within the interval $(11.1 - 32.8)$.

The other main finding in Hedström et al. that will be considered here is stated as follows: "A significant interaction between two genetic risk factors, carriage of human leukocyte antigen DRB1*15 and absence of human leukocyte antigen A*02 was observered among smokers whereas such an interaction was absent among non-smokers." It is natural to interpret this as follows: the data set was split into two groups, one consisting of the non-smokers, and one consisting of the smokers. For each of these groups, the following model (possible with control variables added) was estimated:

$$\begin{aligned} ln\left(\frac{P(MS)}{1 - P(MS)}\right) =& \beta_0 + \beta_1 HLA\_DRB1\_15 \\ &+ \beta_2 HLA\_A\_02 \\ &+ \beta_3 HLA\_DRB1\_15 \times HLA\_A\_02 \end{aligned} \tag{28}$$

The estimate of the coefficient $\hat{\beta}_3$ of the interaction $HLA\_DRB1\_15 \times HLA\_A\_02$ will not be biased due to omission of variables. That is because the variables that are included in the interaction term are included in the model (see the last paragraph in subsection 3.2). The variables HLA_DRB1_15 and HLA_A_02, i.e., the main effects, may well be biased due to omission of variables, as previously discussed.

The results of the regression of equation 28 performed on the group of non-smokers and on the group of smokers are given in tables 11 and 12 respectively.

	$\hat{\beta}_i, i = 1, \ldots, 3$	St. Err.	P-value	$e^{\hat{\beta}_i}$
HLA_DRB1_15	1.155	0.393	0.003	3.174
HLA_A_02	0.285	0.385	0.459	1.330
$HLA\_DRB1\_15 \times HLA\_A\_02$	0.162	0.543	0.765	1.176

Table 11: Estimate of the effect of the two genetic factors including their interaction term on the risk of getting multiple sclerosis. Performed on the group of non-smokers.

	$\hat{\beta}_i, i = 1, \ldots, 3$	St. Err.	P-value	$e^{\hat{\beta}_i}$
HLA_DRB1_15	1.119	0.247	<0.001	3.061
HLA_A_02	0.454	0.244	0.062	1.575
$HLA\_DRB1\_15 \times HLA\_A\_02$	0.515	0.349	0.140	1.674

Table 12: Estimate of the effect of the two genetic factors including their interaction term on the risk of getting multiple sclerosis. Performed on the group of smokers.

The data yield no significant interaction between carriage of Human Leukocyte Antigen DRB1*15 and absence of Human Leukocyte Antigen A*02 among non-smokers (P-value = 0.765), which means that part of Hedström et al. 's result is reproduced. Model 28 is closer to reproducing a significant interaction between HLA_DRB1_15 and HLA_A_02 among smokers (P-value = 0.140).

The lack of statistical strength of these analyses can be further illustrated by considering the confidence intervals of the odds ratios

$$\widehat{OR} = e^{\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3} = e^{\hat{\beta}_1} e^{\hat{\beta}_2} e^{\hat{\beta}_3} \tag{29}$$

computed for the group of non-smokers and the group of smokers respectively. The estimated odds ratios are

$$\widehat{OR}_{non-smokers} = 3.174 \times 1.330 \times 1.176 = 4.96 \tag{30}$$

and

$$\widehat{OR}_{smokers} = 3.061 \times 1.575 \times 1.674 = 8.07. \tag{31}$$

The confidence intervals can be computed using formulas 24, 25 and 27. The estimated standard errors are given by

$$\begin{aligned} &SE_{non-smokers}(\beta_1+\beta_2+\beta_3) \\ &= \sqrt{1.334+0.081+0.026+0.340-0.271-0.066} = 1.20 \end{aligned} \tag{32}$$

and

$$\begin{aligned} &SE_{smokers}(\beta_1+\beta_2+\beta_3) \\ &= \sqrt{1.252+0.206+0.265+0.557-0.814-0.326} = 1.07. \end{aligned} \tag{33}$$

The 95% confidence interval

$$\left(e^{(\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3-1.96SE)}, e^{(\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3+1.96SE)}\right) \tag{34}$$

is therefore approximately given by

$$\begin{cases} \left(\frac{4.96}{10.5}, 4.96\times 10.5\right) = (0.47, 52.1) & \text{in the non-smoker case} \\ \left(\frac{8.07}{8.14}, 8.07\times 8.14\right) = (0.99, 65.7) & \text{in the smoker case.} \end{cases}$$

The point estimate of the effect of having both genetic risk factors among non-smokers, 4.96, is within the confidence interval $(0.99, 65.7)$ for the effect of having both genetic risk factors among smokers. The point estimate of the effect of having both genetic risk factors among smokers, 8.07 is within the confidence interval $(0.47, 52.1)$ for the effect of having both genetic risk factors among non-smokers.

It is furthermore notable that both the confidence intervals include 1, so there can be concluded no effect of both genetic risk factors neither among non-smokers, nor among smokers at a 95% level of significance if we split up the data and include the interaction term.

Instead of splitting the data set into the two groups of non-smokers and smokers and estimate equation 28 for each of these groups, one can keep the whole data set and estimate the following equation:

$$\begin{aligned} ln\left(\frac{P(MS)}{1-P(MS)}\right) =& \beta_0+\beta_1 Smoker \\ &+\beta_2 HLA\_DRB1\_15 \\ &+\beta_3 HLA\_A\_02 \\ &+\beta_4 HLA\_DRB1\_15\times HLA\_A\_02 \\ &+\beta_5 HLA\_DRB1\_15\times Smoker \\ &+\beta_6 HLA\_A\_02\times Smoker \\ &+\beta_7 HLA\_DRB1\_15\times HLA\_A\_02\times Smoker, \end{aligned} \tag{35}$$

where the notation $P(Y)$ is defined in equation 4.

There is a close relationship between the two approaches. The reason why is that if one uses the first method, one estimates the dependent variable MS given each of the values that the variable $Smoker$ can take, namelig 0 and 1, while if one uses the second method, one estimates MS given the events generated by the variable $Smoker$. These are the same conditions. Moreover, if for each variable that is included when using the first method one includes exactly the same variables and in addition the interaction between $Smoker$ and these variables (and no other variables) when using the second method, one will obtain identical point estimates and confidence intervals. For instance, the estimated coefficient 0.162 and standard error 0.543 of the interaction term among non-smokers can be recognized as coefficient β_4 and its standard error in model 35 (see table 13). The coefficient of the interaction term among smokers, 0.515, can be recognized as the sum of the coefficients β_4 and β_7 in model 35, and the standard error of this coefficient, 0.349, can be computed as the standard error of the sum $\beta_4 + \beta_7$.

	$\hat{\beta}_i, i = 1, \ldots, 7$	St. Err.	P-value	$e^{\hat{\beta}_i}$
$Smoker$	0.852	0.333	0.011	2.344
HLA_DRB1_15	1.155	0.393	0.003	3.174
HLA_A_02	0.285	0.385	0.459	1.330
$HLA\_DRB1\_15 \times HLA\_A\_02$	0.162	0.543	0.765	1.176
$HLA\_DRB1\_15 \times Smoker$	-0.036	0.464	0.938	0.964
$HLA\_A\_02 \times Smoker$	0.169	0.456	0.711	1.184
$HLA\_DRB1\_15 \times HLA\_A\_02 \times Smoker$	0.353	0.646	0.585	1.423

Table 13: The result of a regression designed to estimate the effect of the interactions between smoking and the genetic risk factors.

It may be striking how insigificant the estimates are of most of the coefficients in table 13. One explanation is that model 35 includes terms that are highly correlated. Removing some of these will yield more significant estimates. Consider the following model:

$$\begin{aligned} ln\left(\frac{P(MS)}{1-P(MS)}\right) =& \beta_0 + \beta_1 Smoker \\ &+ \beta_2 HLA\_DRB1\_15 \\ &+ \beta_3 HLA\_A\_02 \\ &+ \beta_4 HLA\_DRB1\_15 \times HLA\_A\_02 \times Smoker. \end{aligned} \quad (36)$$

The results of the regression of equation 36 are shown in table 14. Note that the estimate of the coefficient of the second order interaction term now holds a 10% level of significance.

As for the interpretation of the second order interaction term and its estimated coefficient, note first that $HLA\_DRB1\_15 \times HLA\_A\_02 \times Smoker$

	$\hat{\beta}_i, i = 1, \ldots, 4$	St. Err.	P-value	$e^{\hat{\beta}_i}$
$Smoker$	0.884	0.173	<0.001	2.421
HLA_DRB1_15	1.158	0.176	<0.001	3.182
HLA_A_02	0.435	0.175	0.013	1.546
$HLA\_DRB1\_15 \times HLA\_A\_02 \times Smoker$	0.502	0.282	0.075	1.652

Table 14: The results of the regression of a model of interaction between smoking and genetic risk factors. First order interaction factors are omitted.

has the same interpretation in equation 35 as it has in equation 36, as it is the same variable. The estimated coeficients, however, are different due to different statistical strength and, in general, different bias. But the estimated value of the coefficient of the second order interaction term in equation 36, 0.502, cannot be biased due to omission of variables. That is because all the variables that constitute the interaction term are included as controls in model 36 (see similar comments above, in particular the last paragraph of subsection 3.2). For the same reason, the estimated coefficient of the second order interaction term in model 35, 0.353, cannot biased due to omission of variables, but this estimate has a greater standard error.

Recall that the estimated coefficient of the interaction between the genetic factors HLA_DRB1_15 and HLA_A_02 among smokers, 0.515, equals the sum of the coefficients β_4 and β_7 in model 35. Furthermore, the estimated value of the coefficient of the second order interaction term in equation 36, 0.502, is the estimate of β_7, given that $\beta_4 = \beta_5 = \beta_6 = 0$. Therefore, 0.502 is the estimate of the interaction term among smokers, assuming there is no interaction between the two genetic risk factors among non-smokers (which is reasonable considering the results summarized in table 11), and assuming there are no first order interactions between either of the genetic risk factors and smoking (which we concluded already in section 3.2).

4 Discussion

It could be argued that table 9 summarizes the main results of this study. Perhaps most striking is that the data with a high degree of significance indicate that smoking among men increases the risk of getting multiple sclerosis more than fivefold, whereas smoking among women increases that risk only about half as much, and that snuffing seems to approximately halve the risk of getting multiple sclerosis. Again, it should be pointed out that the data do not necessarily imply causal effects. In case of snuffing, one may suspect that it represents something that is protective, and not that snuffing in itself causes lower probability of getting multiple sclerosis.

There may be important effects that are measured via snuffing, or smoking, thus giving a biased estimate of the direct effect. Social factors like

income, education and social status come to mind. It may be a serious drawback to the analysis that the data do not allow for adjusting for such factors. The same comment goes for most environmental factors. In theory, the genetic factors could also be linked to omitted factors, thus distorting the estimate of any direct effects the genetic factors may have on getting multiple sclerosis. What the true mechanisms may be is undetermined be the present study, and may represent topics for future research.

It is presumed that in the case of a multiple sclerosis patient (i.e., a case), it has been registered whether the factors represented by the explanatory variables were present immediately prior to the onset of the disease. As for the genetic data, that is obvious, for a person's genes are constant over time. The same goes for the person's gender. In case of e.g. smoking, this is more of an issue. For the purposes of predicting the onset of a chronic illness, it is of interest whether one was a smoker prior to the onset of the disease. If the patient started smoking after the onset of the disease, any causality must be reverse, from getting multiple sclerosis to becoming a smoker. One problem with the data may be that the individuals may not remember when he or she started to smoke, or that getting the habit of smoking was not really a binary event. The patient may have started out as a party-smoker, and only later have become a habitual smoker. The transition may be blurry and difficult to recall. The same type of comments go for passive smoking and snuffing.

Even if a multiple scerosis patient started out smoking before he or she got the disease, it is a possibility that the smoking did not contribute to cause multiple scerosis, but rather that the person's susceptibility to getting multiple sclerosis contributed to cause the person to start to smoke. A combination of the two is also possible. A regression will pick up the assosiation, but will not by itself determine in what direction(s) any possible causation goes.

Another problem is that it is not obvious that the smoking-data are directly comparable between cases and controls, especially since it ignores possible cumulative effects or effects of dose. For instance, a case may have been smoking lightly for a few years before getting multiple scerosis at 30, and a control may be someone at 60 who has been smoking heavily for more than 40 years. One might believe that the difference in cumulative exposure should somehow matter, but that is not reflected in the data.

The question whether one has been exposed to passive smoking can be ambiguous. It may be obvious to some respondents filling in the questionaire that if one is a smoker and one has never been exposed to passive smoking, one should still be classified as a passive smoker. But the questionaire does not actually ask the individual to fill in whether he or she has been a smoker when answering the passive smoking question. As for the registry of whether the individual has had an Epstein Barr virus infection, it may have been difficult for people to recall whether they have had it at all, and in case of

multiple scerosis patients, whether they had it prior to the onset of their disease.

It is not entirely clear what would be meant by adjusting for age in this context. The basic reason is that the regression is supposed to say something about the risk of getting multiple sclerosis as a function of risk factors. The context indicates that we are interested in the risk of getting multiple sclerosis sometime during the entire life span. A variable *Age* would indicate something qualitatively different, namely the risk of getting multiple sclerosis sometime before or at *Age*. Therefore, the variable *Age* should not be included in the regression, unless we are interested in the risk of getting multiple sclerosis sometime before or at *Age* (and not in the lifetime risk).

The inclusion of the variable *Age* as a control variable may be problematic for yet another reason. Since age and genetic risk factors must be considered to be completely unrelated, any omitted variable bias because of omission of the variable *Age* must be the via environmental factors, i.e., smoking, snuffing, passive smoking or Epstein Barr virus infection. But in case of multiple sclerosis patients, it is environmental exposure prior to the onset of the disease that is relevant. So if the environmental information says anything about the age of the patient, it must be via the mechanisms by which cases were included in the study.

A regression with *Age at onset of the disease* (i.e. the age at which the disease was diagnosed, which is not exactly the same), *Type of multiple sclerosis*, or *The patient's Expanded Disability Status Score (EDSS) score* as explanatory variable is not possible. The reason is that any registered information about these characteristics imply with certainty that the individual has multiple sclerosis. The data items *the age at onset of the disease*, *Type of multiple sclerosis* and *The patient's EDSS score* represent a specification of the variable MS, and are candiates to replace MS as the dependent variable. It is, however, reasonable to start out with an analysis of MS, and possibly proceed to analyses of the alternative dependent variables at a later stage. One could argue that if one is primarily interested in the genetic factors, gender and smoking, and their possible interaction in causing multiple sclerosis, it must surely be more important to have data on the amount and duration of smoking than to try to assess the possible effect of smoking as a binary variable on refined dependent variables.

References

[1] A. Alonso, M. A. Hernan: *Temporal trends in the incidence of multiple sclerosis: a systematic review*, Neurology 2008: 71(2); 129-135 (2008)

[2] B. Brynedal, K. Duvefelt, G. Jonasdottir, I. M. Roos, E. Akesson, J. Palmgren et al.: *HLA-A confers an HLA-DRB1 independent influence on the risk of multiple sclerosis*, PloS.ONE. 2007: 2(7); e664 (2007)

[3] A. K. Hedström, E. Sundqvist, M. Bäärnhielm, N. Nordin, J. Hillert, I. Kockum et al.: *Smoking and two human leukocyte antigen genes interact to increase the risk for multiple sclerosis*, BRAIN 2011: 134; 653-664 (2011)

[4] A. K. Hedström, M. Bäärnhielm, T. Olsson, L. Alfredsson.: *Tobacco smoking, but not Swedish snuff usage, increases the risk of multiple sclerosis*, Neurology 2009: 73; 696-701 (2009)

[5] D. W. Hosmer, S. Lemeshow: *Applied Logistic Regression*, New York: John Wiley & Sons. 1989 (1989)

[6] M. R. Lincoln, A. Monpetit, M. Z. Cader, J. Saarela, D. A. Dyment, M. Tislar et al.: *A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis*, Nature Genetics 2005: 37(10); 1108-1112 (2005)

[7] R. A. Linker, B. C. Kieseier, R. Gold: *Identification and development of new therapeutics for multiple sclerosis*, Trends in Pharmacological Sciences 2008: 29(11); 558-565 (2008)

[8] H. F. McFarland, R. Martin: *Multiple sclerosis: a complicated picture of autoimmunity*, Nat. Immunol. 2007: 8(9); 913-919 (2007)

[9] Web site: *https://lagunita.stanford.edu/c4x/HumanitiesScience /StatLearning/asset/classification.pdf*. Accessed on September 18, 2015.