Multiple Sclerosis: Exploring the Limits of Genetic and Environmental Susceptibility

Goodin DS* [1], Khankhanian P [2], Gourraud PA [1,3,4], Vince N [3,4]

1. Department of Neurology, University of California, San Francisco & the San Francisco VA Medical Center, San Francisco, CA, USA

2. Center for Neuro-engineering and Therapeutics, University of Pennsylvania, Philadelphia, PA, USA

3. Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France.

4. Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France.

Address for Correspondence:

Douglas S. Goodin, MD
Department of Neurology
University of California, San Francisco
UCSF MS Center
675 Nelson Rising Lane, Suite #221D
San Francisco, CA 94158
Phone: (415) 514 2464
Fax: (415) 514 2470
E mail: douglas.goodin@ucsf.edu

Author Contributions

DSG: Conceptualized and led the project, analyzed and assisted data interpretation and the writing of analytic software, developed the conceptual-framework, and wrote the original draft of the manuscript.

PK: Assisted in data interpretation and review of the manuscript.

PA: Assisted in data interpretation and review of the manuscript.

NV: Assisted data interpretation, and the review of the manuscript.

Abstract

OBJECTIVE: To explore the nature of genetic and environmental susceptibility to multiple sclerosis (MS) and to define the limits of this nature based on the statistical uncertainties regarding the various epidemiological observations that have been made.

BACKGROUND: Certain parameters of MS-epidemiology can be directly observed (e.g., the risk of MS-recurrence in siblings and twins of an MS proband, the proportion of *women* among MS patients, the population-prevalence of MS, and the time-dependent changes in the sex-ratio). By contrast, other parameters can only be inferred from the observable parameters (e.g., the proportion of the population that is genetically susceptible, the proportion of *women* among susceptible individuals, the probability that a susceptible individual will experience an environment sufficient to cause MS given their genotype, and if they do, the probability that they will ultimately develop the disease).

DESIGN/METHODS: The "genetically-susceptible" subset (*G*) of the population (*Z*) is defined to include everyone with <u>any</u> non-zero life-time chance of developing MS. For the observed parameters, plausible ranges are assigned values such that they include either the 95% confidence intervals or the estimated parameter-ranges for each observation. By contrast, for the non-observed parameters, the ranges are assigned such that they cover any plausible value for each parameter. Using both a *Cross-sectional Model* and a *Longitudinal Model*, together with established parameter relationships, we explore iteratively trillions of potential parameter combinations and determine those combinations (solutions) that are constrained by the observed parameter ranges.

RESULTS: Using both *Models*, under all circumstances, genetic-susceptibitly is limited to to less than 52% of the population. If solutions are excluded, in which susceptible *women* either comprise less than 18% of susceptible individuals or have a penetrance more than 20 times that in susceptible *men*, then both *Models* agree that genetic-susceptibility is limited to less than 12% of the population. Consequently, a large number (and likely the large majority) of individuals have no chance whatsoever of developing MS, regardless of their environmental exposure. Also, although the penetrance of MS in susceptible *women* is greater than it is in *men* and, for most solutions, susceptible *men* outnumber susceptible *women*. As expected, the probability that susceptible individuals will develop MS increases with an increased likelihood of individuals experiencing a sufficient environmental exposure. Nevertheless, as this probability approaches 1, these response-curves plateau at <64% for *women* and at <23% for *men*. Finally, under current environmental conditions are such that more than 68% of susceptible individuals are experiencing an environment sufficient to cause MS given their particular genotype.

CONCLUSIONS: The development of clinical MS (in an individual) requires both that they have the appropriate genotype (which is quite rare in the population) and that they have an environmental exposure sufficient to cause MS given their particular genotype (which is currently quite common among susceptible individuals). Nevertheless, even when the necessary genetic and environmental factors, required for MS pathogenesis, co-occur for an individual, this is still insufficient for that person to develop MS. Rather, disease pathogenesis, even in this circumstance, involves an important element of chance.

Author Summary

Certain parameters of MS-epidemiology can be directly observed. These parameters include the risk of MS recurrence in siblings and twins of an MS proband, the proportion of *women* among MS patients, the population-prevalence of MS, and the time-dependent changes in the sex-ratio. By contrast, there are other parameters of MS-epidemiology, which can't be observed, but which must be inferred based on the values of the observable parameters. These parameters include the proportion of the general population (*Z*) that is genetically susceptible to MS, the proportion of *women* among susceptible individuals, the probability that a susceptible individual will experience an environment sufficient to cause MS, and if they do, the likelihood that they will actually develop the MS. We define the subset (*G*) – i.e., the genetically-susceptible subset – to include everyone in (*Z*) who has <u>any</u> non-zero chance of developing MS over their life-time. For the observed parameters, plausible ranges are assigned values such that they include either the 95% confidence intervals or the estimated parameter-ranges for each observation. By contrast, for the non-observed parameters, the ranges are assigned such that they cover any plausible value for each parameter. Then, using both a *Cross-sectional Model* and a *Longitudinal Model*, together with established parameter relationships, we explore iteratively trillions of potential parameter combinations and determine those combinations (solutions) that are constrained by the observed parameter values. The *Cross-sectional Model* requires two assumptions, commonly made in studies of monozygotic twins, to establish certain relationships between the observed and non-observed parameters. By contrast, the *Longitudinal Model* makes neither of these assumptions but, rather, this *Model* utilizes the observed changes in the female-to-male (*F:M*) sex-ratio, which have taken place over the past 4–5 decades, to determine the response curves for susceptible individuals, relating their probability of developing MS to their probability of experiencing an environment sufficient to cause MS. Using both of these analytic *Models*, under all circumstances, genetic-susceptibitly is limited to to less than 52% of the population. If solutions are excluded, in which susceptible *women* either comprise less than 18% of susceptible individuals or have a penetrance greater than 20 times that in susceptible *men*, then both *Models* agree that genetic-susceptibility is limited to less than 12% of the population. Consequently, a large number (and likely the large majority) of individuals have no chance whatsoever of developing MS, regardless of their environmental experiences. Also, the penetrance of MS in susceptible *women* is considerably greater than it is in *men* although, for most solutions, susceptible *men* outnumber susceptible *wom*en. As expected, MS-probability increases with an greater likelihood of a sufficient environmental exposure although these response-curves plateau at <64% for *women* and at <27% for *men*. Finally, the current environmental conditions are such that more than 68% of susceptible individuals, at present, are experiencing an environment sufficient to cause MS given their particular genotype. Consequently, the development of clinical MS (in an individual) requires both that they have the

appropriate genotype (which is quite rare in the population) and that they have an environmental exposure sufficient to cause MS given their individual genotype (which is currently quite common among susceptible individuals). Nevertheless, even when the necessary genetic and environmental factors, required for MS pathogenesis, co-occur for an individual, this is still insufficient for that person to develop MS. Rather, disease pathogenesis, even in this circumstance, involves an important element of chance.

Introduction

Susceptibility to multiple sclerosis (MS) is known to be complex, involving the critical interplay between both environmental events and genetic factors [1-3]. Our previously published analysis regarding the nature of this susceptibility [3] was based on a few basic, well-established, epidemiological parameters of MS, which have been repeatedly observed in populations across Europe and North America. These parameters include the prevalence of MS in a population, the recurrence-risk for MS in siblings and twins of MS-probands, the proportion of *women* among MS patients, and the time-dependent changes in the sex-ratio, which have taken place over the last several decades [3]. For the purpose of this analysis, we defined a "genetically-susceptible" subset (*G*) of the general population (*Z*) to include everyone who has any non-zero chance of developing MS over their life-time. We concluded that any such genetic-susceptibility is confined to only a small proportion of these northern populations (<7.3%) and, thus, that the large majority of individuals in these populations have no chance whatsoever of developing MS, regardless of any environmental conditions that they may experience during their lifetimes [3]. Nevertheless, despite this critical dependence of susceptibility to MS upon the genotype of an individual, we also concluded that certain environmental events were also necessary for MS to develop and that, consequently, both essential genetic factors and essential environmental events are in causal pathway leading to MS [3]. If either of these are missing, MS cannot develop. Finally, we concluded that, even when the necessary and sufficient genetic and environmental determinants were present, the actual development of MS, in this circumstance, depended, in part, upon an element of chance [3].

What this analysis did not undertake, however, was to explicitly explore the limits of these conclusions based upon the statistical uncertainties, which surround each of the various epidemiological observations that have been made. It is the purpose of this study, therefore, to undertake such an exploration using both the confidence intervals and estimated ranges for the different basic epidemiological parameters and by incorporating these uncertainties into the governing equations relating these parameters both to each other and to the underlying susceptibility to MS that exists within in a population.

For the purposes of this analysis, we have used, primarily, the data reported from the Canadian Collaborative Project on Genetic Susceptibility to Multiple Sclerosis [4]. The reason for this choice is three-fold. First, this Canadian data is population-based with an initial cohort of 29,478 MS patients who were born between the years 1891 and 1993 [4-6]. Second, this data provides, from the same population, estimates for the recurrence risk in monozygotic (*MZ*) twins, in dizygotic (*DZ*) twins, in non-twin siblings, and for changes in the female-to-male (*F:M*) sex-ratio over time [4-6]. Consequently, this Canadian data is likely among the most complete and the most reliable in the world. And third, these data

come from the same geographic region, which is crucial when considering a disease, which has a markedly variable prevalence in different parts of the world [7].

Methods

1. General Methods

We consider a general population (Z), which is composed of (n) individuals ($k = 1,2,...,n$). In *Table 1*, we define the different parameters used in our analysis and, in *Table 2*, we provide a set of parameter abbreviations, which are used for the purposes of notational simplicity. The term $P(MS)$ represents the life-time probability that a random individual from (Z) will develop MS {i.e., $P(MS) = P(MS \mid Z\}$. This is referred to as the penetrance of MS for the population (Z). The term $P(G)$ represents the probability that a random member of (Z) is also a member of the (G) subset – i.e., $P(G) = P(G \mid Z)$. In turn, the subset (G) subset is defined to consist of <u>all</u> individuals (i.e., unique genotypes) within (Z) who have a non-zero chance of developing MS under some (unspecified and not necessarily realized) environmental conditions, regardless of how small that chance might be. Consequently, each of the ($m \leq n$) individuals in the (G) subset ($i = 1,2,...,m$) has a unique genotype (G_i) such that:

$$\forall\, G_i \in G\text{:}\ P(MS \mid G_i) = x_i > 0$$

and thus: $$P(MS \mid G) > 0$$

Individuals, who do not belong to the (G) subset, belong to the mutually exclusive subset (G–), which consists of <u>all</u> individuals who have no chance, whatsoever, of getting MS, regardless of their environmental experiences. Consequently, each of these ($m_2 = n - m$) "non-susceptible" individuals ($j = 1,2,...,m_2$) has a unique genotype (G_j), such that:

$$\forall G_j \in G-\text{:}\ P(MS \mid G_j) = 0$$

and thus: $$P(MS \mid G-) = 0$$

We also define the set (X) to be the set of non-zero penetrance values for members of the (G) subset. Provided that the variance of (X) is not equal to zero (i.e., $Var(X) = \sigma_X^2 \neq 0$), the subset ($G$) can be further partitioned into at least two mutually exclusive sub-subsets, ($G1$) and ($G2$), suitably defined, such that the expected penetrance for the sub-subset ($G1$) is greater than that for ($G2$). In the present manuscript, ($G1$) is interchangeable with the set of susceptible *women* {i.e., $(G1) = (F \cap G) = (F,G)$} and ($G2$) is interchangeable with the set of susceptible *men* {i.e., $(G2) = (M \cap G) = (M,G)$}.

The term {$P(E)$} represents the probability that a member of the (G) subset will experience an environmental exposure sufficient to cause MS, given the prevailing environmental conditions of the time

(E_T) – i.e., $P(E) = P(E \mid G,E_T)$. This environmental exposure (E) is partitioned conceptually into three components (E_{pop}, E_{sib}, and E_{twn}) such that: $P(E) = P(E_{pop}) + P(E_{sib}) + P(E_{twn})$. The ($E_{pop}$) environment represents those environmental factors experienced with equal liklihood by everyone in the population (including siblings and twins). The (E_{sib}) environment represents those environmental factors experienced either more or less frequently by siblings (including twins) compared to the general population. Presumably, these (E_{sib}) experiences are due to the familial micro-environment and occur mostly (but not necessarily exclusively) during childhood. The (E_{twn}) environment represents those environmental factors experienced either more or less frequently by *MZ*- and *DZ*-twins compared to both siblings and the general population. Presumably, these (E_{twn}) experiences occur mostly (but not necessarily exclusively) during the intrauterine and early post-natal periods. Despite this conceptual framework, however, the observations from Canada in adopted individuals, in siblings and half-siblings raised together or apart, in conjugal couples, and in brothers and sisters of different birth order, have indicated that MS-risk is not affected by the familial micro-environment but, rather, that the important environmental risks (not considering twins) result from population-wide (i.e., E_{pop}) exposures [8-14]. Thus, these studies, collectively, provide compelling evidence for the absence of any (E_{sib}) environmental impact on MS.

We define (E_i) to be the specific set (or sets) of environmental exposures, which are sufficient to cause MS to develop in the ith susceptible individual. This set of exposures is unspecified as to: 1) how many events are involved; 2) when, during the life of an individual, these events need to occur; and 3) what these events are. This set, of whatever it consists, simply needs to be sufficient to cause MS to develop in the ith susceptible individual. Clearly, in order for MS to develop, (G_i) and (E_i) must occur jointly. If either event occurs without the other, MS cannot develop. Also, if every susceptible individual requires the same set of environmental exposures, then $\{P(E_i) = P(E)\}$. If some individuals can develop MS under <u>any</u> environmental conditions, then $\{P(G_i,E_i) = P(G_i)\}$. And, if some environmental conditions can cause MS to develop in <u>anyone</u>, then $\{P(G_i,E_i) = P(E_i)\}$. Considering both genetic and environmental susceptibility (*as defined above*), therefore:

$$P(MS) = P(MS,G,E)$$

Moreover, presumably, genetic susceptibility is independent of the prevailing environmental conditions of the time (E_T). Therefore, $P(E)$ can be expressed as:

$$P(E) = P(E \mid G,E_T) = P(E,G \mid G,E_T)$$

where: $$P(E,G \mid G,E_T) = \sum_{i=1}^{m} P(E_i,G_i \mid G,E_T)$$

$$= \sum_{i=1}^{m} P(G_i \mid G,E_T) * \{P(E_i \mid G_i,E_T)$$

and, where: $\forall G_i \in G$: $P(G_i \mid G,E_T) = P(G_i \mid G) = 1/m$

A set of related terms $P(MS \mid MZ_{MS})$, $P(MS \mid DZ_{MS})$, and $P(MS \mid S_{MS})$ represent the familial occurrences of MS. The 1st two terms, $P(MS \mid MZ_{MS})$ and $P(MS \mid DZ_{MS})$, represent the conditional life-time probability of developing MS for an individual from either an *MZ*- or a *DZ*-twin-ship, given the fact that their co-twin either has or will develop MS. These probabilities are estimated by the observed proband-wise concordance rate for either *MZ*- or *DZ*-twins [15]. In a similar manner, the term $P(MS \mid S_{MS})$ represents the conditional life-time probability of developing MS in a non-twin sibling (*S*), given the fact that their co-sibling either has or will develop MS. Thus, the term (MZ_{MS}) represents the condition in which an *MZ*-twin has (or will develop) MS, regardless of their co-twin's circumstances. The term (MS, MZ_{MS}) represents the condition where the *MZ*-twins are concordant for MS. Consequently, considering the individual genotypes (G_i) within the (*G*) subset, the term (MZ_{MS}) is defined such that:

$$P(MZ_{MS}) = P(MS \mid MZ)$$

and: $$\forall G_i \in G\text{:}\ P(MZ_{MS} \mid G_i) = P(MS \mid G_i, MZ)$$

In addition, the joint probability of an individual is a member of the (*G*) subset, together with having this condition {i.e., $P(G_i, MZ_{MS})$} is defined as:

$$\forall G_i \in G\text{:}\ P(G_i, MZ_{MS}) = P(G_i, MS \mid MZ)$$

In which case: $$\forall G_i \in G\text{:}\ P(G_i \mid MZ_{MS}) = \frac{P(G_i,\ MZ_{MS})}{P(MZ_{MS})} = \frac{P(MS, G_i \mid MZ)}{P(MS \mid MZ)} = P(G_i \mid MS, MZ)$$

Summing over all susceptible individuals and because, by definition, everyone who has MS must be a member of the (*G*) subset (*above*) therefore:

$$\sum_{i=1}^{m} P(G_i, MZ_{MS}) = P(G, MZ_{MS}) = P(G, MS \mid MZ) = P(MZ_{MS})$$

and: $$\sum_{i=1}^{m} P(G_i \mid MZ_{MS}) = P(G \mid MZ_{MS}) = P(G \mid MS, MZ) = 1$$

Lastly, $P(MS \mid IG_{MS})$, represents the proband-wise concordance rate for *MZ*-twins, adjusted for the fact that concordant *MZ*-twins, in addition to sharing their identical genotypes (*IG*), also share their (E_{twn}) environment (see *Table 1*). Such an adjustment may be necessary because, if these shared environmental experiences contribute to MS in the proband, they could also increase likelihood of MS developing in the co-twin and such a circumstance would inflate our estimate for the likelihood of genetic susceptibility in the population. As discussed previously [3], this adjustment (IG_{MS}) is defined such that:

$$\forall G_i \in G\text{:}\ P(MS \mid G_i, IG_{MS}) = P(MS \mid G_i)$$

and: $$\forall G_i \in G\text{:}\ P(G_i \mid IG_{MS}) = P(G_i \mid MZ_{MS})$$

Effectively, this adjustment equates to a thought-experiment, in which *MZ*-twins are separated at conception and then each twin experiences a random (E_{twn}) environment just like any other member of the general population (*Z*), given the environmental conditions of the time (E_T). Clearly, the term $P(MZ_{MS})$,

because it takes no account of what happens to the co-twin, requires no such adjustment. Thus:

$$P(IG_{MS}) = P(MS \mid IG) = P(MS \mid MZ)$$

We have previously demonstrated [3] that this adjustment – i.e., for converting $P(MS \mid MZ_{MS})$ into $P(MS \mid IG_{MS})$ – is made using an adjustment factor (s_a) such that:

$$s_a = P(MS \mid DZ_{MS}) / P(MS \mid S_{MS})$$

and: $P(MS \mid IG_{MS}) = P(MS \mid MZ_{MS}) / s_a$

From these definitions, we can use estimated values for observable population parameters to deduce the value of the non-observable parameter $P(MS \mid G)$, which represents the conditional life-time probability of developing MS for a member of the (G) subset. This term is referred to as the expected penetrance for the (G) subset. As noted above, by definition, everyone who actually develops MS during their life-time must belong to the (G) subset. Therefore:

$$P(MS,G) = P(MS)$$

From this and from the definition of conditional probability:

$$P(MS \mid G) = \mathrm{P}(MS,G\ ) / P(G) = P(MS) / P(G)$$

Rearrangement of this equation yields:

$$P(G) = P(MS) / P(MS \mid G)$$

This parameter value, once established, can then be used to assess the nature of MS pathogenesis. For example, if: $P(G) = 1$, then anyone can get MS under the right environmental circumstances. By contrast, if: $P(G) < 1$, only certain individuals in the general population (Z) have any possibility of getting MS. In this circumstance, MS must be a "genetic" disorder (i.e., unless a person has the appropriate genotype, they have no chance, whatsoever, of getting MS, regardless of their environmental exposure). Importantly, if MS is "genetic" in this sense, this has no bearing whether disease pathogenesis also requires the co-occurrence of specific environmental events.

In this study, two *Models* are used to estimate the values of various unknown epidemiological parameters of MS. The first *Model* takes a cross-sectional approach, in which deductions are made from "current" epidemiological data during a single time-period (E_T). This will be referred to as the *Cross-sectional Model*. The second *Model* takes a longitudinal approach, in which deductions are made both from "current" epidemiological data and from the changes in the female-to-male (*F:M*) sex ratio that have taken place over the past 4–5 decades [5]. This will be referred to as the *Longitudinal Model*. These two *Models* are independent of each other although both incorporate into the analysis many of the same observed and non-observed epidemiological parameters, which are important for MS pathogenesis. The *Cross-sectional Model* derives theoretical relationships between different epidemiologic parameters, but it also makes two assumptions regarding *MZ*-twin data to establish these relationships (*see below; Methods*

#3). These two assumptions are also commonly made by other studies, which analyze *MZ*-twin-data, and each has observational data to support them [16-18]. Nevertheless, both for *Assertions A–D* and for *Equations 4a* and *4b* (*see below*), these conditions need to be assumed. By contrast, the *Longitudinal Model* makes no assumptions to estimate ranges for the non-observed parameters and it also requires a smaller number of parameters, which need to have ranges assigned to them.

For both *Models*, the first step is to assign acceptable ranges for the value of certain "observed" parameters (e.g., twin-concordance rates, the population prevalence of MS, or the proportion of *women* among MS patients). These ranges are assigned such that they include either their calculated 95% confidence intervals or their observed parameter ranges. The second step is to assign acceptable ranges for the "non-observed" parameters (e.g., the proportion of susceptible persons in the population or the proportion of *women* among susceptible individuals). These ranges are assigned such that they cover the entire "plausible" range for each such parameter. In both *Models*, a "substitution" analysis is undertaken to determine those parameter combinations (i.e., solutions) that fit within the constraints set by the acceptable ranges for the observed parameters. These solutions are then used to assess their implications about the basis of genetic and environmental susceptibility to MS in the population. For each *Model*, the total number of parameter combinations interrogated in this manner was $\sim 10^{10}$.

2. Establishing Plausible Ranges for Parameter Values

<u>*Observed Parameter Values*</u>

The observed proband-wise concordance rate for MS in *MZ*-twins in Canada [6] is 25.3%. The 95% confidence interval (*CI*) for $P(MS \mid MZ_{MS})$ is calculated from an exact binomial test [19] as:

$$CI = 0.253 \pm 0.074 = (0.18 - 0.33)$$

The estimates, from different studies, for the current proportion of *women* among the MS patients in North America ranges between 66% and 76% [3]. For our *Cross-sectional Model*, we chose to use the slightly expanded range of:

$$0.66 \leq P(F \mid MS) < 0.78$$

The observed proportion of *women* among the MS patients in Canada was 69% during *Time-Period #1* (1941–1945) and 76% during *Time-Period #2* (1976–1980) [5]. In this study, although the authors do not report the actual numbers of individuals in each 5-year epoch, they do report that these 5-year samples averaged 2,400 individuals per epoch. Also, the authors graphically present the 95% *CIs* for the *F:M* sex ratio during each of these 5-year epochs in their *Figure* [5]. Estimating that the number of individuals in both *Time-Periods #1* and *#2* is ~2,000, and using the exact binomial test, for our

Longitudinal Model, we estimate that for $P(F \mid MS)$:

during *Time-Period #1* (1941–1945): $CI = (0.67 - 0.71)$

during *Time-Period #2* (1976–1980): $CI = (0.74 - 0.78)$

Both of these 95% *CIs* exceed those presented in the manuscript *Figure* [5]. We represent parameter values at the two different *Time-Periods* by subscripts 1 or 2 {e.g., $P(MS)_2$=$P(MS)$ at *Time-Period #2*}.

The observed proband-wise concordance rates for MS in female and male *MZ*-twins in Canada [6] were 34% and 6.7%, respectively. The using the exact binomial test [19], the *CIs* for these observations are:

For: $P(MS \mid F,MZ_{MS})_2$: $CI = (0.24 - 0.44)$

For: $P(MS \mid M,MZ_{MS})_2$: $CI = (0.014 - 0.18)$

The 95% *CI* for the difference in twin concordance between *men* and *women* is calculated as:

$$CI = (\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

In which case: $CI = (0.15 - 0.40)$

According to the 2010 Canadian census [20], the proportion of *women* among the general Canadian population (*Z*) was 50.4%. Thus, *men* and *women* comprise essentially equal proportions of this population and, therefore, by definition:

$$P(F,G) = P(F \cap G) \leq P(F) = 0.5$$

and: $$P(M,G) = P(M \cap G) \leq P(M) = 0.5$$

Previously, we used three independent methods (*based on observation*) to estimate the value of *P*(*MS*) [3]. The first method relied on measures of the population prevalence of MS in North America together with the observed age-distribution for MS-onset, the second method considered the age-specific prevalence of MS in the age-band of (45–54 years), and the third method considered a population-based multiple-cause-death study from British Columbia, which reported the proportion of death certificates that mention MS. The range of values supported, collectively, by these methods was: $0.0025 \leq P(MS) \leq 0.0046$. Nevertheless, for the purposes of the present analysis, we expanded the possible range for this parameter to include:

$$0.001 \leq P(MS) \leq 0.006$$

<u>*Non-Observed Parameter Values*</u>

In addition to the observed parameter values (*above*), we also estimated the values of 7 other parameters $P(G)$; $p = P(F \mid \mathrm{G})$; $x_1 = P(MS \mid F,IG_{MS})$; and the ratios: $C = P(MS)_1/P(MS)_2$;

$s_a = P(MS \mid MZ_{MS}) \,/ P(MS \mid IG_{MS})$; $r = P(MS \mid F,MZ_{MS}) / P(MS \mid F,G)$; and:
$s = P(MS \mid M,MZ_{MS}) / P(MS \mid M,G)$

In several cases, there are constraints on the values that these non-observed parameters can take. For example, because we have defined of one "exposure" unit (*see Methods #4; below*) such that the second time point has greater exposure than the first, we are assuming that *P*(*MS*) is increasing with time for both *men* and *women* such that:

$$C < 1$$

This assumption is in accord with several studies that have found that the prevalence of MS is increasing, especially among *women* [5,21-27]. If this is correct, the ratio (*C*) is constrained on theoretical grounds [3] such that:

$$C < P(M \mid MS)_2 / P(M \mid MS)_1 = 0.238/0.313 = 0.76$$

Naturally, by the same theoretical reasoning [3], it must also be constrained such that:

$$C < P(F \mid MS)_2 / P(F \mid MS)_1 = 0.762/0.687 = 1.11$$

However, only the second of these conditions satisfies all three, so that the upper bound for *C* is 0.76 in this circumstance. Moreover, if ($C < 0.25$) then there must have been a greater than 4-fold increase in *P*(*MS*) for Canada, which has taken place over a 35–40 year-interval. This seems implausible. Therefore, we conclude that, if *P*(*MS*) is increasing in both *men* and *women*, then:

$$0.25 \leq C < 0.76$$

We also consider the possibility that *P*(*MS*) might be decreasing with time in both *men* and *women*. In this circumstance, considering the actual observations [5,21-27], it seems unlikely that any such decrease could be more than 50% so that, in this case:

$$1.11 < C \leq 2$$

The circumstance, in which the *P*(*MS*) in *men* is decreasing over time while, at the same time, it is increasing in *women*, would require that:

$$0.76 < C < 1.11$$

Because the *Equations* governing the different relationships for *men* and *women* in the *Longitudinal Model* are independent of each other (*see Methods #4*), these are the same as those considered above for an increasing MS prevalence in *women* and a decreasing MS prevalence in *men*. Moreover, if the environment actually had such a mixed interaction with gender, this would indicate that the environmental events sufficient to cause MS in *men* <u>must</u> be different from those sufficient to cause MS in *women*. Finally, because the proportion of *women* with MS is increasing with time [5,21-27], there are no circumstances under which the *P*(*MS*) in *women* could be decreasing while, in *men*, it was increasing.

For analyzing the impact of using different *Time-Points* from the Canadian *Sex Ratio* data [5], we used the reported sex ratio change, which had occurred over each time-interval being considered. The upper limit for *C* was then determined in the same manner as that presented *above*. The lower limit for *C* in each case was determined such that this ratio was halved for every 20 years in the past.

For circumstances, in which *P*(*MS*) is decreasing over time in *men* or *women* or both (even for a short time interval), this requires the subscripts (1) and (2) to be interchanged in the *Equations*, both *above* and those in the development of the *Longitudinal Model* (*below*). This is necessary so that the "environmental units" are always defined in such a way that an increasing exposure is positive.

The parameters *P*(*G*) and *P*(*F* | *G*) are unconstrained and, thus, we considered:

$$0.001 \leq P(G) \leq 1 \text{ and: } 0.001 \leq P(F \mid G) < 1$$

Using the definitions provided in *Table 2*, the parameters (x_1) and (x_2) are related such that:

$$x = px_1 + (1-p)x_2 \qquad \textit{Equation 1}$$

This relationship requires no assumptions and (x_1) and (x_2) must <u>always</u> satisfy *Equation #1*, regardless of which *Model* is employed in the analysis [3]. Also, because more penetrant genotypes are enriched in the (G,MS), (F,G,MS) & (M,G,MS) subsets [3] compared, respectively, to the subsets of susceptible individuals (G), (F,G) & (M,G), it must be the case that:

$$x \leq x' \leq P(MS \mid MZ_{MS})$$

$$x_1 \leq x_1' \leq P(MS \mid F, MZ_{MS})$$

$$x_2 \leq x_2' \leq P(MS \mid M, MZ_{MS})$$

Using the observed constraints (*above*) on $P(MS \mid MZ_{MS})$, $P(MS \mid F, MZ_{MS})$ & $P(MS \mid M, MZ_{MS})$ therefore, we considered the plausible ranges for (x), (x_1) & (x_2) to be:

$$0.001 \leq x \leq 0.33$$

$$0.001 \leq x_1 \leq 0.44$$

$$\text{and: } 0.001 \leq x_2 < 0.18$$

Moreover, the observed data [3,6] indicate that:

$$x_1' = 5.1 * x_2'$$

This adjusted penetrance for susceptible *women* with an MZ_{MS} twin {i.e., $P(MS \mid F,IG_{MS})$} is, thus, substantially greater ($p < 0.001$) than that for the comparable subset of susceptible *men* {i.e., $P(MS \mid M,IG_{MS})$}. Because the enrichment is greater for more penetrant genotypes [3], we expect that the difference between (x_1) and (x_2) will generally be less than this, in which case:

$$x_1 < 5.1 * x_2$$

Despite this expectation, however, we also considered more extreme cases, in which *women* are considerably more enriched than anticipated in the $(F \cap MS)$ population compared to the $(F \cap G)$ population. Thus, we considered two other possible scenarios – either the penetrance imbalance is unconstrained or it is constrained such that:

$$x_1 \leq 20 * x_2$$

At the maximum value (*above*) of: $\{P(F \mid MS) = 0.78\}$, this constraint requires that:

$$P(F \mid G) > 0.15$$

At smaller values for $P(F \mid MS)$, this lower limit becomes even lower [3].

Finally, as discussed previously (*see Methods #1*), we estimate the impact (s_a) of the shared (E_{twn}) environment for *MZ*-twins [3] as:

$$s_a = P(MS \mid DZ_{MS}) / P(MS \mid S_{MS}) \geq 1$$

and: $$P(MS \mid IG_{MS}) = P(MS \mid MZ_{MS}) / s_a$$

In the Canadian data [6], the likelihood of developing MS was found to be greater for a *DZ*-twin of an MS-proband compared to that for a non-twin sibling of an MS-proband. From this observation, the value for (s_a) can be estimated [3,6] to be:

$$s_a = P(MS \mid DZ_{MS}) / P(MS \mid S_{MS}) = 0.054/0.029 = 1.86$$

Moreover, this estimate is approximately the same for both *men* and *women* [3]. Thus, these observations from Canada suggest that sharing the (E_{twn}) environment with an MS-proband markedly increases the likelihood of the co-twin developing MS. Nevertheless, it is possible that this difference may be over-estimated by the Canadian data [6]. It is also possible that the impact of the (E_{twn}) environment is underestimated by these data although, in any event [3], it must be the case that:

$$P(MS \mid IG_{MS}) > P(MS \mid S_{MS}) = 0.029 = P(MS \mid MZ_{MS})/8.7$$

Therefore, for the purpose of both *Models* (when this parameter was included in the substitution analysis), we considered the plausible range to be:

$$1 \leq s_a \leq 8.6$$

However, because from the Canadian data, the estimate [3] for the ratio:

$$s_a = P(MS \mid DZ_{MS}) / P(MS \mid S_{MS})$$

is generally greater than that reported in other similar studies [6,28-34], we also considered, separately, the more restrictive circumstance in which:

$$1 \leq s_a \leq 1.9$$

3. *Cross-sectional Model:*

For the simplicity of notation (*Table 2*) in developing the *Cross-sectional Model*, we define the following abbreviations:

$x = P(MS \mid G)$; $x' = P(MS \mid IG_{MS})$; $x_1 = P(MS \mid G1)$; $x'_1 = P(MS \mid G1, IG_{MS})$;

$x_2 = P(MS \mid G2)$; $x'_2 = P(MS \mid G2, IG_{MS})$; σ_X^2 = variance of the set (X); $p = P(G1 \mid G)$;

$r = x'_1/x_1$; and: $s = x'_2/x_2$

Assertions: ***A.*** $\forall G_i \in G$: $P(G_i, MS \mid MZ) = P(G_i, MS)$

and: $\forall G_i \in G$: $P(G_i \mid MZ_{MS}) = P(G_i \mid MS)$

B. $P(MZ_{MS}) = P(IG_{MS}) = P(MS)$

C. $x = (x'/2) \pm (\sqrt{(x')^2 - 4\sigma_X^2})/2$

D. $0 \leq \sigma_X^2 < (x'/2)^2$ and: $\sigma_X^2 = x(x' - x)$

Definitions and Assumptions: The subsets (G) and (G–) have already been defined *(see Methods #1)* and as noted there: $\forall G_i \in G$: $x_i = P(MS \mid G_i)$

Thus, (x_i) represents the MS-penetrance for the ith individual in the (G) subset and it is unique to the ith individual, regardless of whether it is identical (in value) to that of another individual. Also, as discussed in *Methods #1*, we defined the set (X) to include the penetrance value for every member of the (G) subset. Thus: $X = \{x_i\}$

Treating (x_i) as a random variable, representing any element of the set (X), it follows that:

$$E(x_i) = P(MS \mid G) = x$$

and: $E(x_i - x)^2 = E(x_i^2) - x^2 = \sigma_X^2$

As noted earlier, if $\{\sigma_X^2 \neq 0\}$, then the subset (G) can be partitioned into two mutually exclusive sub-subsets, ($G1$) and ($G2$), suitably defined, such that:

$$x_1 > x_2$$

Moreover, as defined earlier – *see Methods #1*: $P(G_i, MZ_{MS}) = P(G_i, MS \mid MZ)$

Therefore: $\forall G_i \in G$: $P(G_i \mid MZ, MS) = P(G_i, MS \mid MZ)/P(MS \mid MZ)$

$$= P(G_i, MZ_{MS})/P(MZ_{MS}) = P(G_i \mid MZ_{MS})$$

We now specify our two assumptions:

Assumption #1

Because *MZ*-twinning is generally thought to be non-hereditary [16-18], we assume that every genotype (i.e., person) in the general population (*Z*) has the same chance, *a priori*, of having an *MZ*-twin (i.e., *MZ*-status is independent of genotype). Thus, we assume that:

$$\forall G_k \in Z\text{:}\ \ P(MZ \mid G_k) = P(MZ)$$

Because this relationship applies to every genotype in (*Z*), therefore, also:

$$\forall G_i \in G\text{:}\ \ P(MZ \mid G_i) = P(MZ)$$

Even for circumstances, in which *MZ*-twinning is only rarely familial [18] it will be the case that:

$$\forall G_k \in Z\text{:}\ \ P(MZ \mid G_k) \approx P(MZ)$$

Finally, even if rare cases of *MZ*-twinning are thought to be familial [18], the exact relationship (*second Equation above*) still follows if we assume that those genetic factors, which are associated with *MZ*-twining, are independent of those genetic factors associated with MS.

Assumption #2

The penetrance of MS for an *MZ*-twin, whose co-twin is of unknown status, is assumed to be the same as if that genotype had occurred without having an *MZ*-twin (i.e., the penetrance of MS for each genotype is independent of *MZ*-status). This assumption translates to assuming that the impact of experiencing any particular (E_{twn}) environment together with an *MZ*-twin is the same as the impact of experiencing the same (E_{twn}) environment alone. Alternatively, it translates to the testable hypothesis that the mere fact of having an *MZ*-twin does not alter the (E_{twn}) environment in such a way that MS becomes more likely in both twins. Following the definitions relating to (MZ_{MS}) – *see Methods #1*, we are here assuming that:

$$\forall G_i \in G\text{:}\ \ P(MS \mid G_i, MZ) = P(MS \mid G_i)$$

Proof of Assertion A:

From *Assumption #1*, it follows that:

$$\forall G_k \in Z\text{:}\quad P(G_k, MZ) = P(G_k) * P(MZ \mid G_k) = P(G_k) * P(MZ)$$

and therefore: $\forall G_k \in Z\text{:}\ \ P(G_k, MZ)/P(MZ) = P(G_k \mid MZ) = P(G_k)$

Consequently, also: $\forall G_i \in G\text{:}\ \ P(G_i) = P(G_i \mid MZ)$

From this conclusion, from the definitions relating to (MZ_{MS}) – *see Methods #1* – and from *Assumption #2*, it follows that:

$$\forall G_i \in G\text{:}\ \ P(G_i, MZ_{MS}) = P(MS, G_i \mid MZ) = P(G_i \mid MZ) * P(MS \mid G_i, MZ)$$
$$= P(G_i) * P(MS \mid G_i) = P(MS, G_i)$$

and:

$$P(MZ_{MS}) = P(MS \mid MZ) = \sum_{i=1}^{m} P(G_i \mid MZ) * P(MS \mid G_i, MZ)$$
$$= \sum_{i=1}^{m} P(G_i) * P(MS \mid G_i) = P(MS)$$

Therefore, from the definitions relating to (MZ_{MS}) – *see Methods #1* – and from these two equivalences, we conclude that:

$$\forall G_i \in G: \; P(G_i \mid MS, MZ) = P(G_i \mid MZ_{MS}) = \frac{P(G_i, MZ_{MS})}{P(MZ_{MS})} = \frac{P(G_i, MS)}{P(MS)} = P(G_i \mid MS)$$

Proof of Assertion B:

From *Assertion A* and from the definition of (IG_{MS}) – see *Methods #1* – it follows that:

$$P(IG_{MS}) = P(MZ_{MS}) = P(MS)$$

Proof of Assertion C:

As noted above, the variance (σ_X^2) of the set (X) can be expressed as:

$$\sigma_X^2 = E(x_i - x)^2 = E(x_i^2) - x^2$$

Where, because all genotypes in a population are unique:

$$\forall G_i \in G: \; P(G_i \mid G) = 1/m$$

Also:

$$P(G) = (m/n)$$

$$E(x_i) = \sum_{i=1}^{m} (x_i) * (1/m)$$

$$E(x_i^2) = \sum_{i=1}^{m} (x_i)^2 * (1/m) = x^2 + \sigma_X^2$$

From the definitions of $P(G)$ and (IG_{MS}) – *see Methods #1* – it directly follows that:

$$P(MS \mid G_i, IG_{MS}) = P(MS \mid G_i, G, IG_{MS}) = P(MS \mid G_i, G) = P(MS \mid G_i) = x_i$$

1. Consequently, the probability $P(MS, G_i \mid G, IG_{MS})$ can be re-expressed as:

$$P(MS, G_i \mid G, IG_{MS}) = P(G_i \mid G, IG_{MS}) * P(MS \mid G_i, G, IG_{MS})$$
$$= P(G_i \mid G, IG_{MS}) * P(MS \mid G_i) = P(G_i \mid G, IG_{MS}) * (x_i)$$

2. In turn, the term $P(G_i \mid G, IG_{MS})$ can be re-expressed as:

$$P(G_i \mid G, IG_{MS}) = P(G_i \mid G, MS) = P(G_i, G, MS)/P(MS, G)$$
$$= P(MS \mid G_i, G) * P(G_i, G)/P(MS, G)$$

$$= (x_i) * P(G_i \mid G) / P(MS \mid G) = (x_i) * (1/m)/x$$

Combining *1 & 2 (above*) yields:

$$P(MS, G_i \mid G, IG_{MS}) = \{(x_i) * (1/m)/x\} * (x_i) = (x_i)^2 * (1/m)/x$$

However: $x' = P(MS \mid G, IG_{MS}) = P(MS, G \mid G, IG_{MS}) = \sum_{i-1}^{m} P(MS, G_i \mid G, IG_{MS})$

Where: $\sum_{i-1}^{m} P(MS, G_i \mid G, IG_{MS}) = \sum_{i=1}^{m} (x_i)^2 (1/m)/x = E(x_i^2)/x$

Therefore: $x' = (x^2 + \sigma_X^2)/x = x + \sigma_X^2/x$ (*Equation 2*)

Equation #2 can be rearranged to yield a quadratic equation in (*x*) such that:

$$x^2 - (x')x + \sigma_X^2 = 0$$

This quadratic can be solved to yield:

$$x = (x'/2) \pm \sqrt{(x')^2 - 4\sigma_X^2} / 2 \qquad \text{(Equation 3)}$$

Proof of Assertion D:

Equation #3 has real solutions only for the range of:

$$0 \leq \sigma_X^2 \leq (x')^2/4 = (x'/2)^2$$

Equation #2 can also be re-arranged to yield:

$$\sigma_X^2 = x(x' - x)$$

Assertion C Solutions

Equation 2 has two solutions – the so-called <u>*Upper Solution*</u>, {as: $(x \longrightarrow x')$} and the <u>*Lower Solution*</u> {as: $(x \longrightarrow 0)$} depending upon the value of the (±) sign of *Equation #3*. The <u>*Upper Solution*</u> represents the gradual transition from a distribution, in which everyone has a penetrance of (*x'*), to a bimodal distribution, in which half of the (*G*)-subset has a penetrance of (*x'*) and the other half has a penetrance of zero [3]. However, by definition: $(\forall\, x_i \in X\colon\ x_i > 0)$. Therefore, the <u>*Upper Solution*</u> is constrained such that:

$$x'/2 < x \leq x'$$

The <u>*Lower Solution*</u> represents the gradual transition from the bimodal distribution described above to an increasingly extreme and asymmetric distribution [3]. The <u>*Lower Solution*</u> is constrained both by the requirement that when: $(\sigma_X^2 = 0)$ then: $(x = x')$ and by the same considerations as for the <u>*Upper*</u>

Solution. Therefore, the *Lower Solution* is constrained such that:

$$0 < x < x'/2$$

Moreover, regardless of whether the *Upper* or *Lower Solution* pertains, the values that (x_1) and (x_2) can take are further constrained [3]. Thus, using the notation presented above and in *Table 2*:

$$x_1 = \frac{x + \sqrt{x^2 - \{1 + (r/s)(1-p)/p)\}\{x^2 - xx'(1-p)/s\}}}{p + (r/s)(1-p)} \qquad (\textit{Equation 4a})$$

and:

$$x_2 = \frac{x - \sqrt{x^2 - \{1 + (s/r)(p/(1-p)\}\{x^2 - xx'\ p/r\}}}{(1-p) + (s/r)p} \qquad (\textit{Equation 4b})$$

Although the derivation of *Equations 4a & 4b* does not depend upon any of the conclusions of *Assertions A–D* (*above*), nevertheless, one step in the derivation [3] requires the equivalence:

$$P(G1 \mid MS,MZ) = P(G1 \mid MS)$$

or, in the context of the present manuscript, that:

$$P(F \mid MS,MZ) = P(F \mid MS)$$

Consequently, *Equations 4a & 4b* make the same two assumptions.

For this analysis we also assume that either the set (G) by itself or, considered separately, the sets ($G1$) and ($G2$), conform to the *Upper Solution* [3]. In this circumstance, on theoretical grounds [3], it must be the case that either:

$$1 \le x'/x < 2 \quad \text{or both:} \quad 1 \le r = x_1'/x_1 < 2 \ ; \quad \text{and:} \quad 1 \le s = x_2'/x_2 < 2$$

Using a "substitution" analysis, we wrote a computer program, which incorporated the above acceptable ranges (*see Methods #1*) for the parameters {$P(G)$; $P(MS \mid MZ_{MS})$; $p=P(F \mid G,)$; r; s; $P(MS)$; $P(F \mid MS)$; and s_a}, into the governing equations and determined those combinations (i.e., solutions) that fit within the constraints imposed by the "observed parameter" ranges.

4. *Longitudinal Model*:

Following standard survival analysis methods [35], we define the cumulative survival {$S(u)$} and failure {$F(u)$} functions where: $F(u) = 1 - S(u)$, in addition to the hazard-rate functions `$h(u)$} and $g(u)$ for developing MS at different exposure-levels (u) in "susceptible" *men* and *women*, respectively. Furthermore, we define the exposure (u) as the odds that a member of the (G) subset receives an

environmental exposure sufficient to cause MS so that:

$$u = P(E)/[1 - P(E)]$$

The hazard-rate functions may or may not be proportional. However, if they are proportional, then: $g(u) = R * h(u)$; where *R* represents the proportionality constant. If these hazard functions are not proportional, then *men* and *women* need to be analyzed separately using their own scales and the parameter (λ) is meaningless (*see below*). Presumably, such a circumstance (i.e., non-proportional hazard) indicates that *men* and *women* require different environmental conditions for MS to develop [3].

We then transform the exposure scale (for *men*) from (u) units into (a) units. To do this, we define *H(u)* to be the definite integral of the hazard-function $h(u)$ from a ($u = t$) exposure-level to a ($u = 0$) exposure-level and then define (a) units as:

$$\forall\ (t)\ in\ the\ domain\ of\ u: \qquad a = H(u) = \int_0^t h(u)du$$

Because these new units are arbitrary, "one unit" of environmental exposure in *men* can be defined as the difference in exposure between any two *Time Periods* (e.g., a_1 and a_2) such that:

$$a_2 - a_1 = 1$$

This presumes that: ($a_2 > a_1$). Otherwise, to ensure that "one unit" of exposure is positive, the subscripts (1) and (2) need to be interchanged both here and in all the Equations relating to this *Model*.

In a similar manner, and with the same caveats, we can transform exposure (for *women*) into a different scale of "apparent" exposure units (a^{app}) such that:

$$\forall\ (t)\ in\ the\ domain\ of\ u: \qquad a^{app} = G(u) = \int_0^t g(u)du$$

This scale is termed apparent because it may not be the same scale as in *men*. On this scale, "one unit" of environmental exposure (regardless of proportionality) is defined as:

$$a_2^{app} - a_1^{app} = 1$$

Initially, we restrict our considerations of hazard to *men*. Nevertheless, because the scale to which *men* and *women* are assigned is arbitrary, these considerations apply equally and analogously to *women*. Thus, using the standard definitions [35] that:

$$f(u) = d[F(u)]/du = -d[S(u)]/du$$

and: $$h(u) = f(u)/S(u)$$

a typical derivation from survival analysis methods [35] demonstrates that:

because: $$h(u)du = -d\{S(u)\}/S(u)$$

therefore: $$\ln[S(u)] = -\int h(u)du$$

Consequently, the cumulative survival is exponentially related to the integral of the underlying hazard function. Because we are measuring exposure as the odds that a member of the (G) subset receives an environmental exposure sufficient to cause MS, given the environmental conditions at the time, by

definition, when: $\{P(E) = 0\}$, no member of (G) can develop MS {i.e., $S(0) = 1$} and, therefore, at the exposure level ($u = 0$):

$$\ln[S(u)] = \ln(1) = 0 = -\int h(u)du$$

In this circumstance:

$$\forall\ (u = t): \qquad a = H(u) = \int_0^t h(u)du = \int h(u)du$$

so that: $da = h(u)du$

Thus, for *men*: $\ln[S(a)] = -\int da = -a$

and, for *women*: $\ln[S(u)] = -\int g(u)du$

so that: $\ln\left[S(a^{app})\right] = -\int d(a^{app}) = -\,a^{app}$

and when the hazards are proportional: $\ln\left[S(a^{app})\right] = -\,a^{app} = -Ra$

Thus, for *men*: $S(a) = e^{-a}\ \ and\text{:}\ \ F(a) = 1 - e^{-a}$

and, for *women*: $S(a^{app}) = e^{-a^{app}}\ \ and\text{:}\ \ F(a^{app}) = 1 - e^{-a^{app}}$

As noted earlier, we use subscripts (*1*) and (*2*), respectively, to denote the failure probabilities and the parameter values at the 1st and 2nd *Time-Periods*. However, unlike true survival (where everyone dies given sufficient time), the probability of developing MS may not approach 100%, as $P(E)$ approaches 1. In addition, the limiting value for the probability of developing MS in susceptible *men* (***c***) need not be the same as that in susceptible *women* (***d***). Nevertheless, the threshold environmental exposure (at which MS becomes possible) must occur at: $P(E) = 0$; for, at least, one of these two sub-subsets – provided that this exposure level is possible [3]. The difference in threshold (λ) is defined as the difference between the threshold in *women* (λ_w) and the threshold in *men* (λ_m) – i.e., ($\lambda = \lambda_w - \lambda_m$). Thus, if the threshold in *women* is greater than the threshold in *men*, (λ) will be positive.

Also, in true survival, both the clock and the risk of death begins immediately at time-zero and continues indefinitely, so that the cumulative probability of death always increases with time. By contrast, here, it may be that the prevailing environmental conditions during some period of time (E_T) are such that $P(E) = 0$ – even for a quite extended period of time (e.g., centuries or millennia). Moreover, unlike the cumulative probability of death, here, exposure can vary in any direction with time, depending upon the environmental conditions and, thus, the cumulative probability of failure (i.e., developing MS) can vary positively or negatively with time.

From these definitions, the failure probability for susceptible *women* (*Zw*) and *men* (*Zm*) at the *1st Time Period* is:

$$F(a^{app})_1 = Zw_1 = P(MS,E \mid G,F)_1 = \boldsymbol{d} * [1 - e^{-a_1^{app}}] \qquad (women)$$

and: $F(a)_1 = Zm_1 = P(MS,E \mid G,M)_1 = \boldsymbol{c} * [1 - e^{-a_1}]$ (*men*)

Following the definitions of one "exposure-unit" (*see above*), at the 2nd *Time Period*, these *Equations* become:

$$F(a^{app})_2 = Zw_2 = P(MS,E \mid G,F)_2 = \boldsymbol{d} * \{1 - e^{-(a_1^{app}+1)}\} \qquad (women)$$

and: $F(a)_2 = Zm_2 = P(MS,E \mid G,M)_2 = \boldsymbol{c} * \{1 - e^{-(a_1+1)}\}$ (*men*)

For both *men* and *women*, the observations at *Time-Periods* (*1*) and (*2*) represent two points on their respective response curves. Because these curves are exponential, any two observation points will define the entire curve. Therefore, we can use the observed (*F:M*) sex-ratio change in Canada between any two *Time-Periods* [5] and, thereby, construct each of these response curves [3]. The values for Zw_2, Zm_2, Zw_1, Zm_1, (***c***), (***d***), $P(E \mid F)$, $P(E \mid M)$, and (λ) can then be determined [3] as:

$$Zw_2 = P(MS,E \mid G,F)_2 = P(MS \mid G,F)_2 = P(F,MS)_2/P(G,F)$$

$$Zm_2 = P(MS,E \mid G,M)_2 = P(MS \mid G,M)_2 = P(M,MS)_2/P(G,M)$$

$$Zw_1 = P(MS,E \mid G,F)_1 = \{P(F,MS)_2/P(F,MS)_1\} * C^2 * Zw_2$$

$$Zm_1 = P(MS,E \mid G,M)_1 = \{P(M,MS)_2/P(M,MS)_1\} * C^2 * Zm_2$$

$$\boldsymbol{c} = (Zm_2) * \{1 - [P(M,MS)_1/P(M,MS)_2] * e^{-1}\}/(1 - e^{-1})$$

$$\boldsymbol{d} = (Zw_2) * \{1 - [P(F,MS)_1/P(F,MS)_2] * e^{-1}\}/(1 - e^{-1})$$

$$P(E \mid F,G)_2 = (Zw_2)/\boldsymbol{d}$$

$$P(E \mid M,G)_2 = (Zm_2)/\boldsymbol{c}$$

and: $\lambda = \ln(1 - Zw_2/\boldsymbol{d}) - \ln(1 - Zm_2/\boldsymbol{c})$

Notably, the value for each the above parameters is independent of whether the hazards are assumed to be proportional and, moreover, the values for $P(F \mid MS)_2$, $P(F \mid MS)_1$, and $P(MS)_2$ have been directly or indirectly observed [3,5]. In addition, even if the hazards are proportional, none of these estimates depends upon the value of the proportionality constant (*R*). =However, in the case of (λ), if the hazards are not proportional, *men* and *women* cannot be plotted on the same graph and (λ) is meaningless [3]. Also, the values of (Zw_1), (Zw_2), (Zm_1), (Zm_2), $P(E \mid F)_2$, $P(E \mid M)_2$, (***c***) and (***d***) are, not surprisingly, only related to the circumstances of either *men* or *women*, considered separately. Using a "substitution" analysis, we wrote a computer program, which incorporated the acceptable parameter ranges (*see*

Methods #2; above) for the six parameters {$P(G)$; $P(MS|F,G)_2$; $P(F|G,)$; $P(MS)_2$; $P(F|MS)_1$; and C}, into the above governing equations and determined those combinations (i.e., solutions) that fit within the constraints imposed by the observed parameter ranges. We also wrote a separate program to incorporate the additional parameters of {$P(MS|MZ)_2$; r; s ; and s_a} into this analysis so that these ranges could also be estimated from the observed data. In this latter circumstance, we loosened our constraints on the values of (r) and (s) such that:

$$1 \leq r = x_1'/x_1 \leq 30 \;\; ; \;\; \text{and:} \;\; 1 \leq s = x_2'/x_2 \leq 30$$

In addition, because (*as noted above*), any two points on an exponential curve define the curve completely, we also undertook an analysis using this *Model* to estimate the range of possible parameter values using each of the possible "two-point" observations from the Canadian study [5]. In this analysis the possible ranges for: $P(G)$; $P(MS|MZ)_2$; $P(MS|F,G)_2$; $p=P(F|G,)$; $P(MS)_2$; $P(F|MS)_1$; C; r; s; and s_a, were incorporated into the above governing equations and, again, determined those combinations (solutions) that fit within the constraints imposed by the observed parameter ranges (*see Methods #2*). In this analysis, we estimated the range of average parameter values for the 9 consecutive 5-year estimates, for the 8 consecutive 10-year estimates, for the 7 consecutive 15-year estimates, and so forth up to the single estimate for the entire study period [5]. Separately, we also estimated the same parameter values for the observations between the two epochs (1966–1970 & 1971–1975), when there was, apparently, a reversal in the general trend that the proportion of *women* was increasing over time (reversing the subscripts as described in the *Methods #4*).

Finally, we estimated the parameter ranges based on the possibility that, contrary to the empirical evidence [5,21-27], $P(MS)$ might be declining over time for *men* or *women* or both.

Results

1. *Cross-sectional Model*

Assuming that the subset (G) conforms to the <u>*Upper Solution*</u> of the *Cross-sectional Model*, and using *Assertion C* (*above*) the range of values for the parameters $P(G)$, $P(MS|G)$ and $P(F,G)$ were:

$$0.003 \leq P(G) < 0.53$$

$$0.01 < P(MS|G) \leq 0.33$$

$$0.002 \leq P(F,G) < 0.25$$

If we consider the more restricted range of ($s_a < 1.8$) for the impact of sharing the (E_{twn}) environment with an *MZ*-twin, then:

$$0.003 \leq P(G) < 0.12$$

$$0.05 < P(MS \mid G) \leq 0.33$$

$$0.003 \leq P(F,G) < 0.07$$

Assuming that the subset (G) does not conform to the <u>*Upper Solution*</u>, but that, considered separately and again using *Assertion C*, both of the subsets (F,G) & (M,G) do, the range of values for the parameter $P(G)$ is:

$$0.003 < P(G) < 0.7$$

$$0.001 \leq P(MS \mid G) < 0.15$$

$$0.001 < P(F,G) < 0.32$$

In addition, the range for (s_a) is restricted such that: ($s_a < 1.8$):

$$0.003 < P(G) \leq 0.53$$

$$0.001 \leq P(MS \mid G) \leq 0.15$$

$$0.002 \leq P(F,G) < 0.07$$

2. Longitudinal Model

Using the *Longitudinal Model*, with only the constraints imposed by the observed parameters (*see Methods #2*), and assuming an increasing value of $P(MS)$ over time in both *men* and *women*, the range of possible ranges for the parameters $P(G)$, $P(MS \mid G)$ and $P(F,G)$ were:

$$0.001 < P(G) < 0.52$$

$$0.002 \leq P(MS \mid G) \leq 0.2$$

$$0.002 \leq P(F,G) < 0.015$$

Also, using the *Longitudinal Model,* incorporating the additional parameters: $P(MS \mid MZ)_2$; r ; s; and s_a – *see Methods #4* , we found that the solution space for (r) was restricted such that*:*

$$1 \leq r < 1.9$$

By contrast, (s) was unrestricted with solutions including ($s = 30$). Thus, using the *Longitudinal Model*, the subset (F,G) must conform to the <u>*Upper Solution*</u> of the *Cross-sectional Model*.

In addition, in this analysis, the possible range for the adjustment of sharing the (E_{twn}) environment with an *MZ*-twin who ultimately develops MS was restricted such that:

$$1 \leq s_a < 1.8$$

Thus, by the *Longitudinal Model*, there were no solutions with an (E_{twn})-adjustment substantially greater than that observed in Canada [5].

Again, excluding values of ($p < 0.18$), the estimate for $P(G)$ becomes:

$$0.001 < P(G) < 0.12$$

This same range for $P(G)$ is also estimated if we apply the constraint that:

$$x_1 \leq 20 * x_2$$

This last constraint also restricts the value of (p) such that ($p \geq 0.17$) and, thus, these two constraints are effectively equivalent.

As anticipated, for all solutions ($\lambda > 0$) so that, if the hazards are proportional, *men* have a lower environmental threshold for developing MS than *women* [3].

The proportion of *women a*mong susceptible individuals {i.e., $P(F \mid G)$} varied widely:

$$0.01 < P(F \mid G) < 0.75$$

although, for 85% of the solutions:

$$P(F \mid G) < 0.5$$

Moreover, when: $P(F \mid G) \geq 0.5$; then: $P(G) \leq 0.04$.

In addition, the maximum likelihood of developing MS in *men* (***c***) and *women* (***d***), given the necessary genotype, together with a sufficient environmental exposure appropriate for that individual, are:

$$0.30 \leq \boldsymbol{d} = P(MS \mid E,F,G) \leq 0.64$$

$$0.001 \leq \boldsymbol{c} = P(MS \mid E,M,G) < 0.18$$

$$2 < (\boldsymbol{d}/\boldsymbol{c}) < 708$$

and: If: $(\boldsymbol{d}/\boldsymbol{c}) \leq 5$: then: $P(G) \leq 0.01$

Finally, the proportion of genetically susceptible individuals who currently experience environmental conditions sufficient to cause MS given their individual genotype {i.e., $P(E \mid G)_2$} is:

$$0.68 \leq P(E \mid F,G)_2 = Zw_2/\boldsymbol{d} \leq 0.86$$

$$0.70 \leq P(E \mid M,G)_2 = Zm_2/\boldsymbol{c} \leq 0.99$$

The *Figure* shows these parameter estimates derived from analyses using different intervals of time for the Canadian data regarding the change in the *Sex-Ratio* over time. As can be appreciated from the *Figu*re, these estimates are quite similar (and consistent) for all of the potential intervals of analysis. This conclusion includes analysis of the single interval between the epochs of (1966–1970) and (1971–1975), during which there was a reversal in the general trend towards an increasing proportion of *women* over time (*Figure*).

Finally, for the analysis assuming the value of $P(MS)$ has been decreasing over time in both *men* and *women*, and applying the same acceptable ranges given in the *Methods #2*:

$$\boldsymbol{c} < 1$$

$$\boldsymbol{d} < 1$$

$$P(G) < 0.52$$

$$P(F, G) < 0.032$$

$$P(G \mid F) \leq 0.66$$

$$00.35 \leq P(E \mid F,G)_2 \leq 0.76$$

$$00.43 \leq P(E \mid M,G)_2 \leq 0.86$$

$$\lambda < 0$$

$$1.02 < (\boldsymbol{d}/\boldsymbol{c}) < 221$$

and: If: $(\boldsymbol{d}/\boldsymbol{c}) \leq 5$: then: $P(G) \leq 0.07$

Discussion

The present analysis provides considerable insight to the nature of susceptibility to MS. Firstly, both of our *Models* substantially support each other. For example, considering the intersection of both *Models,* the supported range for *P*(*G*) is:

$$0.003 < P(G) < 0.52$$

This range is also supported regardless of circumstances, even including the unlikely possibility that, despite several observations to the contrary [5,21-27], *P*(*MS*) has been declining over the four decades span represented by the data from Canada. Thus, in any circumstance, a substantial proportion of the general population (>48%) must be impervious to getting MS, regardless of their environmental experiences. Moreover, if either an extreme gender imbalance within the (*G*) subset is excluded as implausible {i.e., $P(F \mid G) < 0.18$}, or an extreme, and unexpected, imbalance in penetrance is present {i.e., $P(MS \mid F,G) > 20 * P(MS \mid M,G)$} then, considering both *Models*, the estimated range for *P*(*G*) becomes:

$$0.003 \leq P(G) < 0.12$$

and an even greater proportion of the general population (>88%) must be impervious to getting MS.

In addition, for every solution using the *Longitudinal Model*, the plateau of the response curve in is higher in *women* than *men* (i.e., $\boldsymbol{d} > \boldsymbol{c}$) and when this difference between genders is five-fold or less, fewer than 7% of the population can possibly be members of the (*G*) subset. In this circumstance, an even greater proportion of the population (>93%) has no chance, whatsoever, of getting MS. In this sense, MS is fundamentally a genetic disease. If a person doesn't have the appropriate genotype, they can't get the disease and a large majority of the population lacks this essential component of MS pathogenesis.

However, MS is also a fundamentally environmental disease. Thus, the fact that $P(MS,E \mid G)$ is changing over time indicates that environmental factors are equally critical. This conclusion is also indicated by observations that *P*(*MS*) is increasing over time in many parts of the world [5,21-27]. It is also suggested by the observations that: $P(MS \mid G,MZ_{MS}) < 0.33 \ll 1$ and that the (E_{twn}) environment

seems to significantly impact the likelihood of developing MS in the future [6,28-34]. Indeed, in our *Longitudinal Model*, which makes none of the assumptions made by our *Cross-sectional Model*, the upper bound for this impact of the (E_{twn}) environment is both substantial and approximately the same as that, which has been observed empirically in Canada [6], and which has also been reported elsewhere [28-34].

In the present analysis, using a much wider range of parameter values than previously [3], the preponderance of *men* among genetically susceptible individuals was less clear than it seemed before. Nevertheless, for 85% of the possible combinations (solutions) identified in this analysis, the proportion of *men* in the subset (G) exceeded that of *women* {i.e., $P(F \mid G) < 0.5$)}. Such a result might be viewed as surprising given the fact that all but one of the 233 genes linked to MS are on autosomal chromosomes [36]. To rationalize this apparent dilemma, we previously designated (G_{ak}) to represent each of the (n) autosomal genotypes ($k = 1,2,...,n$) in the general population (Z). Moreover, as we noted previously [3], in this circumstance, it is certainly possible that:

$$\forall\, G_{ak} \in Z : P(G_{ak} \mid M) = P(G_{ak} \mid F) = 0.5 * P(G_{ak})$$

and, yet, for some specific genotypes to have the characteristic that:

$$(G_{ak}, M) \in G \quad \text{and:} \quad (G_{ak}, F) \notin G$$

Nevertheless, despite MS pathogenesis depending critically upon both the genotype of an individual and their environmental exposure, it is also clear that there must be another factor is at play. As noted above, over the 45–50 year period spanned by the Canadian data, $P(MS)$ seems to have been increasing for both *men* and *women* both in Canada and in many parts of the world [5,21-27]. In this circumstance, more than 36% of susceptible *women* and more than 77% of susceptible *men* will never develop clinical MS, even when the correct genetic background occurs together with an environmental exposure sufficient to cause MS in those individuals [3]. Moreover, these estimates are unchanged (*Figure*), regardless of which time-interval is analyzed during the 45–50 years represented by Canadian data [5]. Some of these individuals, no doubt, will have subclinical disease [e.g., 37-41]. Collectively, from these autopsy studies, the prevalence of "asymptomatic" MS is ~0.1% and this figure is generally supported by several magnetic resonance imaging (MRI) studies of asymptomatic individuals [42]. Although this figure suggests that a substantial proportion of MS can be asymptomatic, this seems unlikely to account for the large difference between our observations and the expected 100% occurrence of MS in individuals who are both genetically susceptible and, in addition, experience an environment sufficient to cause MS given their particular genotype. This conclusion is especially compelling for *men*, where their response curve plateaus at a significantly lower level than that in *women* (i.e., $\boldsymbol{d} > \boldsymbol{c}$) and also, where *men* account for only 16% of the asymptomatic individuals detected by MRI [42]. Consequently, it

seems that there must be a stochastic factor in MS pathogenesis [3], which determines, in part, who gets the disease and who does not.

During the development of our *Longitudinal Model* of analysis, we observed that when the prevailing environmental conditions of the time (E_T) were such that $\{P(E) = 0\}$, no member of (G) could develop MS. Previously, we considered the possibility that such an environment might not be possible to achieve because some susceptible individuals might be able to develop MS under <u>any</u> environmental conditions – i.e., if these cases were "purely genetic" [3]. Upon further reflection, however, such a possibility seems remote. For example, the first clinical description of MS was published in 1868 by Charcot although earlier pathological descriptions predated this clinical description by ~30 years [43,44]. Perhaps, the earliest described case of MS was that of Saint Lidwina of Schiedam (c. 1396) although the argument for Augustus d'Este (c. 1822) having MS is more unequivocal [43]. Despite the fact that many human afflictions were initially described during the advent of modern medicine in the 19th century, MS is a rather distinctive disorder, and it seems likely that case reports (that would be familiar to us) would have appeared in earlier eras. Moreover, with the onset of the industrial revolution in the late 18th or early 19th century, the environmental conditions began to change substantially (especially for humans). Therefore, both MS as a disease and permissive environmental conditions seem likely to be of relatively recent onset. More importantly, since its original description, MS seems to be changing in character. For example, although considered uncommon initially, ever since Charcot's clinical characterization, MS has become increasingly recognized as a common neurological disease [3,44]. Also, in the 19th century Charcot's triad of limb ataxia, nystagmus (internuclear ophthalmoplegia), and scanning (cerebellar) speech was considered typical whereas, today, while this triad still occurs, such a syndrome is unusual [44-46]. Moreover, in the late 19th and early 20th centuries, the disease was thought to be more (or equally) prevalent in *men* compared to *women* [44,45,47], whereas, today, *women* account for 66–76% of the cases [3]. Also, in many parts of the world, MS is increasing in frequency, particularly among *women* [5,21-27]. Indeed, in Canada, if $P(MS)$ is increasing in both *men* and *women*, it has increased by a minimum of 32% over a 35 to 40 year span [3] – a circumstance which has led to a significant 10% increase in the proportion of *women* among MS patients ($p < 10^{-7}$) over the same time-interval [5].

By contrast, those genetic markers, which are associated with MS, seem to have been present for far greater periods of time. For example, the best established (and strongest) genetic associations with MS are for certain haplotypes within the *HLA* region on the short arm of chromosome 6, including haplotypes such as *DRB1*15:01-DQB1*06:02; DRB1*03:01- DQB1*02:01*; and *A*02:01-C*05:01-B*44:02* [48-52]. Each of these haplotypes is well-represented in diverse human populations around the globe [53] and, thus, these haplotypes must be of ancient origin. Presumably, therefore, the absence of MS prior to the

late 14th (and possibly the early 19th) century, together with the markedly changing nature of MS over the past 200 years, points to a change in environmental conditions as the basis for the recent occurrence of MS as a clinical entity and for the changes in MS pathogenesis, which have taken place over the past two centuries. Consequently, it seems that $\{P(E) = 0\}$ is possible under those environmental circumstances that existed prior to the 14th century and, thus, that "purely genetic" MS does not exist.

In summary, the development of clinical MS (in an individual) requires both that they have the appropriate genotype (which is quite rare in the population) and that they have an environmental exposure sufficient to cause MS given their particular genotype (which is currently quite common among susceptible individuals). Nevertheless, even when the necessary genetic and environmental factors, required for MS pathogenesis, co-occur for an individual, this is still insufficient for that person to develop MS. Rather, disease pathogenesis, even in this circumstance, involves an important element of chance.

Figure. Values for certain epidemiological parameters, estimated from the *Longitudinal Model* using the Canadian data regarding the change in Sex-Ratio, assuming $P(MS)$ is increasing over time in both *men* and *women* [5]. Parameter abbreviations are as indicated in *Table 2*, including:

$$\boldsymbol{d} = P(MS \mid E,G,F); \quad \boldsymbol{c} = P(MS \mid E,G,F)$$

In all cases, the estimates are for circumstances in which:

$$P(MS,E \mid G,F) \leq 20 * P(MS,E \mid G,M)$$

Time-span (in years) indicates the number of years in the interval from which the parameter values are estimated. For example, "5y" represents the average change that occurred between the 9 consecutive 5-year epochs, "10y" represents the average change that occurred between the 8 consecutive 10-year epochs, and so forth. The time-span labeled "Reversal" represents the parameter estimates taken from single change that occurred between the two 5-year epochs (1966–1970 & 1971–1975), when the proportion on *men* increased [5].

References

1. Gourraud PA, Harbo HF, Hauser SL, Baranzini SE. (2012) The genetics of multiple sclerosis: an up-to-date review. *Immunol Rev* 248:87–103.

2. Hofker MH, Fu J, Wijmenga C. (2014) The genome revolution and its role in understanding complex diseases. *Biochim Biophys Acta* 1842:1889-1895.

3. Goodin DS, Khankhanian P, Gourraud PA, Vince N. (2018) Genetic and Environmental Susceptibility to Multiple Sclerosis. *PLoS One* 16(3): e0246157

4. Sadovnick AD, Risch NJ, Ebers GC. Canadian collaborative project on genetic susceptibility to MS, phase 2: rationale and method. Canadian Collaborative Study Group. *Can J Neurol Sci* 1998;25:216–21.4.

5. Orton SM, Herrera BM, Yee IM, et al., and the Canadian Collaborative Study Group. (2006) Sex ratio of multiple sclerosis in Canada: A longitudinal study. *Lancet Neurol*. 5:932–6.

6. Willer CJ, Dyment DA, Rusch NJ, Sadovnick AD, Ebers GC, the Canadian Collaborative Study Group. (2003) Twin concordance and sibling recurrence rates in multiple sclerosis. *Proc Natl Acad Sci U S A.* 100:12877–82.

7. Goodin DS. (2014) The Epidemiology of Multiple Sclerosis: Insights to Disease Pathogenesis. In. *Handbook of Clinical Neurol*ogy Aminoff MJ, Boller F, Swaab DF (eds). Elsevier, London 122:231-266.

8. Bager P, Nielsen NM, Bihrmann K, Frisch M, Wohlfart J, et al. (2006) Sibship characteristics and risk of multiple sclerosis: A nationwide cohort study in Denmark. *Am J Epidemiol* 163:1112–1117.

9. Compston A, Coles A (2002) Multiple sclerosis. *Lancet* 359:1221–31.

10. Dyment DA, Yee IML, Ebers GC, Sadovnick AD, and the Canadian Collaborative Study Group (2006) Multiple sclerosis in step siblings: Recurrence risk and ascertainment. *J Neurol Neurosurg Psychiatry* 77:258–259.

11. Ebers GC, Sadovnick AD, Dyment DA, Yee IM, Willer CJ, et al. (2004) Parent-of-origin effect in multiple sclerosis: observations in half-siblings. *Lancet* 363:1773–1774.

12. Ebers GC, Yee IML, Sadovnick AD, Duquette P, and the Canadian Collaborative Study Group

(2000) Conjugal multiple sclerosis: Population based prevalence and recurrence risks in offspring. *Ann Neurol* 48:927–931.

13. Sadovnick AD, Yee IML, Ebers GC, and the Canadian Collaborative Study Group (2005) Multiple sclerosis and birth order: A longitudinal cohort study. *Lancet Neurol* 4:611–617.

14. Sadovnick AD, Ebers GC, Dyment DA, Risch NJ, and the Canadian Collaborative Study Group (1996) Evidence for genetic basis of multiple sclerosis. *Lancet* 347:1728–1730.

15. Witte JS, Carlin JB, Hopper JL. (1999) Likelihood-Based Approach to Estimating Twin concordance for dichotomous traits. *Genetic Epidemiol*. 16:290–304.

16. Hankins GVD, Saade GR. (2005) Factors influencing twins and zygosity. *Paediatr Perinat Epidemiol* 19(*Suppl 1*):8-9.

17. Hoekstra C, Zhao ZZ, Lambalk CB, Willemsen G, Martin NG, Boomsma DI, Montgomery GW. (2008) Dizygotic twinning. *Hum Reprod Update* 14:37-47.

18. Machin G. (2009) Familial Monozygotic Twinning: A Report of Seven Pedigrees. *Am J Med Genet* 151C:152-154.

19. Newcombe R.G. (1998). Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine*, 17:857–872.

20. Canadian Census (2010) https://www150.statcan.gc.ca/n1/en/pub/89-503-x/2010001/article/11475-eng.pdf?st=WVL9_Ggm

21. Hernán MA, Olek MJ, Ascherio A. (1999) Geographic variation of MS incidence in two prospective studies of US women. *Neurology* 53:1711–1718.

22. Koch-Henriksen N. (1999) The Danish Multiple Sclerosis Registry: a 50-year follow-up. *Mult Scler.* 5:293-296.

23. Celius EG, Vandvik B. (2001) Multiple sclerosis in Oslo, Norway: prevalence on 1 January 1995 and incidence over a 25-year period. *Eur J Neurol* 8:463-469.

24. Barnett MH, William DB, Day S, et al. (2003) Progressive increase in incidence and prevalence of multiple sclerosis in Newcastle, Australia: a 35-year study. *J Neurol Sci* 213:1 – 6.

25. Sarasoja T, Wilkström J, Paltamaa J, et al. (2004) Occurrence of multiple sclerosis in central

Finland: a regional and temporal comparison during 30 years. *Acta Neurol Scand* 110:331-6. O

26. Freedman DM, Mustafa M, and Alavanja MCR. (2000) Mortality from multiple sclerosis and exposure to residential and occupational solar radiation: A case-control study based on death certificates. *Occup Environ Med* 57:418–421.

27. Sundström P, Nyström L, Fosgren L. Incidence (1988-97) and prevalence (1997) of multiple sclerosis in Västerbotten County in northern Sweden. *J Neurol Neurosurg Psychiatry* 2003;74:29-32.

28. French Research Group on Multiple Sclerosis. (1992) Multiple sclerosis in *54* twinships: Concordance rate is independent of zygosity. *Ann Neurol* 32:724-727.

29. Mumford CJ, Wood NW, Kellar-Wood H, Thorpe JW, Miller DH, Compston DA. (1994) The British Isles survey of multiple sclerosis in twins. *Neurology.* 44:11–5.

30. Hansen T, Skytthe A, Stenager E, Petersen HC, Brønnum-Hansen H, Kyvik KO. (2005) Concordance for multiple sclerosis in Danish twins: an update of a nationwide study. *Mult Scler.* 11:504–10.

31. Hansen T, Skytthe A, Stenager E, Petersen HC, Kyvik KO, Brønnum-Hansen H. (2005) Risk for multiple sclerosis in dizygotic and monozygotic twins. *Mult Scler*. 11:500–3.

32. Islam T, Gauderman WJ, Cozen W, et al. (2006) Differential twin concordance for multiple sclerosis by latitude of birthplace. *Ann Neurol* 60: 56–64.

33. Ristori G, Cannoni S, Stazi MA, et al. and the Italian Study Group on Multiple Sclerosis in Twins. (2006) Multiple sclerosis in twins from continental Italy and Sardinia: A Nationwide Study *Ann Neurol* 59:27–34.

34. Kuusisto H, Kaprio J, Kinnunen E, et al. (2008) Concordance and heritability of multiple sclerosis in Finland: Study on a nationwide series of twins. *Eur J Neurol* 15: 1106–1110.

35. Fisher LD, van Belle G. (1993) Biostatistics: A Methodology for the Health Sciences, John Wiley & Sons, New York, pp. 786-829.

36. International Multiple Sclerosis Genetics Consortium. (2019) Multiple sclerosis genomic map

implicates peripheral immune cells and microglia in susceptibility. Science 365 (6460): .doi:10.1126/science.aav7188.

37. Vost A, Wolochow D, Howell D. (1964) Incidence of infarcts of the brain in heart diseases. *J Path Bact* 88:463-470.

38. Georgi VW. (1966) Multiple Sklerose: Pathologisch-Anatomische Befunde multiple Sklerose bei klinisch nicht diagniostizierte Krankbeiten. *Schweiz Med Wochenschr* 20:605-607.

39. Gilbert J, Sadler M. (1983) Unsuspected multiple sclerosis. *Arch Neurol* 40:533-536.

40. Engell T. (1989) A clinical patho-anatomical study of clinically silent multiple sclerosis. *Acta Neurol Scand* 79:428-430.

41. Okuda DT, Mowery EM, Cree BAC, Crabtree EC, Goodin DS, Waubant E, Pelletier D. Asymptomatic spinal cord lesions predict disease progression in radiologically isolated syndrome. *Neurology* 2011;76:686-692.

42. Granberg T, Martola J, Kristoffersen-Wiberg M, Aspelin P, Fredrikson S. Radiologically isolated syndrome--incidental magnetic resonance imaging findings suggestive of multiple sclerosis, a systematic review. *Mult Scler* 2013;19:271-280.

43. Murray TJ. (2004) Multiple sclerosis: The history of a disease. In: Multiple Sclerosis: Diagnosois, Medical Management, and Rehabilitation. Burks JS, and Johnson KP (eds). Demos Medical Publishing. pp. 1-32.

44. Gowers WR. (1888) A manual of diseases of the nervous system. P. Blakeston, Son & Co. Philadelphia, PA, USA. pp. 919-930.

45. Kinnier Wilson SA (1940) Neurology. Edward Arnold & Co. London, pp. 145-178.

46. Compston A, Confavreux C, Lassman H, McDonald I, Miller D, Noseworthy J, Smith K, Wekerly H. McAlpine ‘s Multiple Sclerosis. (2006) Churchill Livingstone, Elsevier pp.287-346

47. Wechsler IS. Statistics of multiple sclerosis. *Arch Neurol Psychiat* 1922;8:59-75.

48. Dyment DA, Herrera BM, Cader Z, Willer CJ, Lincoln MR, Sadovnock AD, Risch N, Ebers GC. Complex interactions among MHC haplotypes in multiple sclerosis: susceptibility and resistance.

Hum Mol Genet 2005;14:2019-2026.

49. Link J, Kockum I, Lorentzen AR, Lie BA, Celius EG, Westerlind H, Schaffer M, Alfredsson L, Olsson T, Brynedal B, Harbo HF, Hillert J. Importance of Human Leukocyte Antigen (HLA) Class I and II Alleles on the Risk of Multiple Sclerosis. *PLoS One* 2012 ;7(5):e36779.

50. Patsopoulos NA, Barcellos LF, Hintzen RQ, Schaefer C, van Diujn CM, Nobel JA, Raj T, IMSGC, ANZgene, Gourraud PA, Stranger BE, Oksenberg J, Olsson T, Taylor BV, Saucer S, Hafler DA, Carrington M, De Jager PL, de Bakkar PIW. (2014) Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. *PLoS Genet* 9(11):e1003926.

51. Chao MJ, Barnardo MC, Lincoln MR, Ramagopalan SV, Herrera BM, Dyment DA, Montpetit A, Sadovnick AD, Knight JC, Ebers GC. HLA class I alleles tag HLA-DRB1*1501 haplotypes for differential risk in multiple sclerosis susceptibility. *Proc Natl Acad Sci USA* 2008;105:13069-74.

52. Multiple Sclerosis Genetics Group. Linkage of the MHC to familial multiple sclerosis suggests genetic heterogeneity. *Hum Molec Genet* 1998;7:1229–1234.

53. Granger L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum Immunol* 2013;74:1313-1320.

Table 1. Definitions for the Groups and Epidemiological Parameters used in the Analysis

Parameter	**Definition**
(Z)	Set of all individuals (i.e., unique genotypes) in the population
$P(MS)$	Expected life-time probability of developing *MS* for a member of (Z)
(G)	Subset of all individuals (genotypes) in (Z) who have <u>any</u> non-zero chance of getting MS under some (unspecified) environmental conditions
$(G-)$	Subset of all individuals (genotypes) in (Z) who have no (zero) chance of getting MS, regardless of environmental conditions
$(G1)$	The subset of "high-penetrance" genotypes in (G): here $(G1) = (F, G)$
$(G2)$	The subset of "low-penetrance" genotypes in (G): here $(G2) = (M, G)$
(M)	Set of all *men* in (Z)
(F)	Set of all *women* in (Z)
(MZ)	Set of all individuals who have a mono-zygotic twin
$P(E)$	Probability of an environmental exposure "sufficient to cause MS" in the subset (G), given the prevailing environmental conditions of the time. $P(E) = P(E \mid G, E_T)$
(E_T)	The prevailing environmental conditions of during a specific *Time-Period* (T)
(E_{pop})	That part of a sufficient environmental exposure shared with equal likelihood by the population generally.
(E_{sib})	That part of a sufficient environmental exposure likely to be shared by siblings (including twins) due to the familial microenvironment – especially, but not necessarily only, during childhood
(E_{twn})	That part of a sufficient environmental exposure likely to be shared by twins (*MZ*- or *DZ*) – especially, but not necessarily only, during the *IU* and early post-natal period.
(E_i)	The environmental exposure sufficient to cause MS in the ith individual in (G)
$P(MS \mid MZ_{MS})$	Expected life-time probability of developing *MS* for an *MZ*-twin whose co-twin either has, or will develop, *MS*
$P(MS \mid DZ_{MS})$	Expected life-time probability of developing *MS* for a *DZ*-twin whose co-twin either has, or will develop, *MS*
$P(MS \mid S_{MS})$	Expected life-time probability of developing *MS* for a non-twin sibling whose co-sibling either has, or will develop, *MS*
$P(MS \mid IG_{MS})$	Value of $P(MS \mid MZ_{MS})$, which has been adjusted to exclude the impact of the similar (E_{twn}) environment of *MZ*-twins
$P(MZ_{MS})$ $P(IG_{MS})$	Expected life-time probability of developing *MS* for <u>any</u> individual from an *MZ*-twin-ship. $P(MZ_{MS}) = P(MS \mid MZ) = P(MS \mid IG) = P(IG_{MS})$

Table 2. Principal Parameter Abbreviations

Parameter	Definition
x_i	$\forall G_i \in G$: $P(MS \mid G_i) = x_i$; -- the penetrance of the (i^{th}) genotype in (G)
X	Set of all penetrance values in the (G) subset -- $X = \{ x_i \}$
x	$= P(MS \mid G)$ -- the expected penetrance for the (G) subset
x'	$= P(MS \mid IG_{MS})$ -- the adjusted *MZ*-twin concordance for MS
σ_X^2	Variance of the set (X) -- $Var\ (X)$
x_1	$= P(MS \mid G1)$ -- the expected penetrance for the ($G1$) subset
x_1'	$= P(MS \mid G1, IG_{MS})$ -- adjusted *MZ*-twin concordance for the ($G1$) subset
x_2	$= P(MS \mid G2)$ -- the expected penetrance for the ($G2$) subset
x_2'	$= P(MS \mid G2, IG_{MS})$ -- adjusted *MZ*-twin concordance for the ($G2$) subset
u	$= P(E) / [1 - P(E)]$ -- Odds of a sufficient environmental exposure
$h(u)$	Hazard-rate function for *men*
$g(u)$	Hazard-rate function for *women*
R	$= g(u) / h(u)$ -- Proportionality constant for hazard (*if proportional*)
Subscripts (1,2)	Indicate Parameter Values at *Time Period #1* or *Time Period #2*
C	$= P(MS)_1 / P(MS)_2$ -- ratio of $P(MS)$ at *Time #1* to that at *Time #2*
p	$= P(G1 \mid G)$ -- the proportion of the (G) subset that is also in ($G1$)
r	$= (x_1' / x_1)$
s	$= (x_2' / x_2)$
Zw	$= P(MS, E \mid G, F)$
Zm	$= P(MS, E \mid G, M)$
$\boldsymbol{c}$	$= P(MS \mid G, E, M)$ -- the limiting value of the exponential curve for *men*
$\boldsymbol{d}$	$= P(MS \mid G, E, F)$ -- the limiting value of the exponential curve for *women*
s_a	$= P(MS \mid DZ_{MS}) / P(MS \mid S_{MS})$ -- adjustment factor necessary because *MZ* & *DZ*-twins share (E_{twn}) -- $P(MS \mid IG_{MS}) / P(MS \mid MZ_{MS}) / s_a$

Time Span	*d*	*c*	*d/c*	*P*(*G*)	*P*(*F*|*G*)	*P*(*E*|*G,F*)	*P*(*E*|*G,M*)
5y	0.27–0.48	0.02–0.17	2.4–19.7	0.01–0.10	0.2–0.5	0.87–0.96	0.89–0.99
10y	0.27–0.52	0.02–0.18	2.4–20.2	0.01–0.10	0.2–0.5	0.80–0.93	0.83–0.98
15y	0.26–0.55	0.01–0.19	2.4–20.7	0.01–0.20	0.1–0.5	0.77–0.90	0.80–0.97
20y	0.26–0.55	0.01–0.19	2.5–21.5	0.01–0.12	0.1–0.5	0.76–0.89	0.81–1.00
25y	0.26–0.59	0.03–0.20	2.1–13.7	0.01–0.09	0.2–0.6	0.72–0.87	0.78–0.99
30y	0.26–0.56	0.03–0.20	2.1–14.1	0.01–0.10	0.2–0.8	0.71–0.84	0.76–0.98
35y	0.27–0.58	0.03–0.21	2.1–14.2	0.01–0.05	0.2–0.6	0.69–0.82	0.72–0.97
40y	0.26–0.61	0.03–0.21	2.1–14.7	0.01–0.05	0.2–0.6	0.70–0.81	0.73–0.97
45y	0.26–0.63	0.01–0.21	2.1–19.1	0.01–0.10	0.2–0.6	0.67–0.78	0.70–0.93
Reversal	0.31–0.48	0.02–0.17	2.7–19.4	0.01–0.10	0.1–0.4	0.87–0.97	0.88–0.98

Figure