Multiple Sclerosis: Exploring the Limits of Genetic and Environmental Susceptibility

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

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. Nantes Université, INSERM, Center for Research in Transplantation and Translational Immunology, UMR 1064, F-44000 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, wrote the analytic software, analyzed and interpreted the data, 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 in data interpretation, and review of the manuscript.

Acknowledgement:

We are especially indebted to John Petkau, PhD, Professor Emeritus, Department of Statistics, University of British Columbia, Canada. Dr. Petkau helped immeasurably with this project; devoting numerous hours of his time to critically reviewing early drafts of this manuscript and providing an invaluable contribution both to the clarity and to the logical development of the mathematical and statistical arguments presented herein.

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 are directly observable (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 female-to-male (*F:M*) *sex-ratio*. By contrast, other parameters can only be inferred from observed 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 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 under *some* environmental conditions. For the observed parameters, acceptable ranges are assigned values such that they always include their 95% confidence intervals. By contrast, for the non-observed parameters, the acceptable ranges are assigned such that they cover the entire "plausible" range 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 (i.e., solutions) that fall within the acceptable range for the observed and non-observed parameters.

RESULTS: Both *Models* and all analyses are consistent and converge to demonstrate that genetic-susceptibitly is limited to 52% or less of the population and to 30% or less of *women*. Consequently, most individuals (particularly *women*) have no chance whatsoever of developing MS, regardless of their environmental exposure. Also, currently, the penetrance of MS in susceptible *women* is greater than it is in *men*. Moreover, as expected, the probability that susceptible individuals will develop MS increases with an increased likelihood of these individuals experiencing an environment sufficient to cause MS, given their genotype. Nevertheless, although it is conceivable that these response-curves plateau at 100% for both *women* and *men*, this possibility requires extreme conditions and seems remote. Rather, at least *men*, seem to plateau well below this level and, if so, it is this difference, rather than any differences in the genetic and environmental determinants of disease, that primarily accounts both for the difference in penetrance between *women* and *men* and for the increasing proportion of *women* among of MS patients worldwide.

CONCLUSIONS: The development of MS (in an individual) requires both that they have an appropriate genotype (which is uncommon in the population) and that they have an environmental exposure sufficient

to cause MS given their individual genotype. Nevertheless, even when the necessary genetic and environmental factors, sufficient for MS pathogenesis, co-occur for an individual, this still insufficient for that person to develop MS. Thus, disease pathogenesis, even in this circumstance, seems not to be deterministic but, rather, to involve 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 female-to-male (*F:M*) *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, in fact, 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, under some environmental circumstances. For the observed parameters, plausible ranges are assigned acceptable values such that they always include their 95% confidence interval. By contrast, for the non-observed parameters, the acceptable ranges are assigned such that they cover the entire "plausible" range 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 (i.e., solutions) that are allowed by the observed and non-observed parameter ranges. The *Cross-sectional Model* makes 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* and the disease prevalence, 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. Both *Models* and all analyses are consistent with each other and converge to demonstrate that genetic-susceptibitly is limited to 52% or less of the population and 30% or less of women. Consequently, most individuals have no chance whatsoever of developing MS, regardless of their environmental experiences. Thus, MS is a genetic disease in the sense that, if an individual does not have the correct genetic makeup, they can't develop the disease. However, the probability that susceptible individuals will develop MS increases with an increased likelihood of these individuals experiencing an environment sufficient to cause MS, given their genotype. Thus, MS is also and environmental disease in the sense that the development of MS (in an individual), in addition to their having an appropriate genotype, requires that they experience an environmental exposure sufficient to cause MS given their individual genotype. Nevertheless, there must be another factor involved in disease pathogenesis because, although it is conceivable that these response-curves plateau at 100% for both *women* and *men*, this possibility requires extreme conditions and seems remote. Rather, at

least *men*, seem to plateau well below this and, if so, it is this difference, rather than differences in the genetic and environmental determinants of disease, that primarily accounts both for the difference in penetrance between *women* and *men* and for the increasing proportion of *women* among of MS patients worldwide. Consequently, even when the necessary genetic and environmental factors, sufficient for MS pathogenesis, co-occur for an individual, this still seems to be insufficient for that person to develop MS. Thus, disease pathogenesis, even in this circumstance, seems not to be deterministic but, rather, to involve 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 individuals with MS, the proportion of *women* among MS patients, and the time-dependent changes in both the female-to-male (*F:M*) *sex-ratio* and the disease prevalence, which have taken place over the last several decades [3]. For 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 the course their lifetime. We concluded that genetic susceptibility, so-defined, is limited to only a small proportion of these northern populations (<7.3%) and, thus, that most 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 the causal pathway leading to MS [3]. If either of these are missing, MS cannot develop. Finally, we concluded that, seemingly, even when the sufficient genetic and environmental determinants were present, the actual development of MS 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 (*CIs*) and "plausible" 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 this analysis, we have used, primarily, the data reported from the Canadian Collaborative Project on Genetic Susceptibility to Multiple Sclerosis [4,5]. The reason for this choice is three-fold. First, this Canadian dataset is a *population-based* sample with an initial cohort of 29,478 MS patients who were born between the years 1891 and 1993 [4-7]. This cohort consists of all MS patients seen in 15 MS Centers scattered throughout the Canadian Provinces [5]. The cohort did not specifically include patients from the Northern Territories [5] although, likely, many of these patients were referred for 2nd opinions to the provincial centers. This study endeavored to include most (or all) of the MS patients in Canada at the

time and, indeed, the authors estimate (from their twin studies) that their ascertainment scheme captured 65-83% of all Canadian MS patients [7]. Importantly, also, this cohort ascertained cases from throughout Canada. Therefore, for the purposes of our analysis, this cohort is assumed to represent a large random sample of the symptomatic Canadian MS population at the time. Second, this dataset provides, from the same population, estimates for the recurrence-risk in monozygotic (*MZ*) twins, in dizygotic (*DZ*) twins, in non-twin siblings (*S*), and for changes in the (*F:M*) *sex-ratio* over time [4-7]. Consequently, this Canadian dataset is likely among the most complete and the most reliable in the world. And third, these data come from a geographic region of similar latitude, which is critical when considering a disease, for which disease prevalence has a marked latitudinal gradient in different parts of the world [8].

Methods

1. General Methods

A. General Model Specifications and Definitions for Genetic Susceptibility to MS

We consider a general population (Z), which is composed of (N) individuals $(k = 1,2, \dots, N)$ who are living under the prevailing environmental conditions during some specific *Time Period* (T) – conditions that are designated, generically, as (E_T). 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 subset (MS) is defined to include *all* individuals within (Z) who either have or will develop MS over the course of their lifetime. The occurrence of (MS) represents the event that an individual, randomly selected from (Z), belongs to this (MS) subset and the term $P(MS \mid E_T)$ represents the probability of this event, given the prevailing environmental conditions of (E_T) – i.e., $P(MS \mid E_T) = P(MS \mid Z, E_T)$. This probability is referred to as the "*penetrance*" of MS for the population (Z) during the *Time Period* (E_T). The occurrence of (G) is defined as the event that an individual, randomly selected from (Z), is a member of the (G) subset. The term $P(G \mid E_T)$ represents the probability of this event given the prevailing environmental conditions of (E_T). In turn, the (G) subset is defined to include *all* individuals (genotypes) within (Z) who have *any* non-zero chance of developing MS under *some* (unspecified and not necessarily realized) environmental conditions, regardless of how small that chance might be. We assume that a person's genotype is independent of the environmental conditions that prevail during (E_T). Therefore:

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

In this circumstance, each of the $(m \leq N)$ individuals in the (G) subset $(i = 1,2, \ldots, m)$ has a unique genotype (G_i). The occurrence of (G_i) represents the event that an individual, randomly selected from (Z), belongs to the (G_i) subset – a subset consisting of only a single individual (i.e., the so-called "i^{th} susceptible individual" or "i^{th} individual") – and the term $\{P(G_i) = 1/N\}$ represents the probability of this event. Therefore, it follows from the definition of the (G) subset that, if <u>*every*</u> relevant environmental condition – *see below* – is possible during some *Time Period* (E_T), then, during this *Time Period*:

$$\forall\, G_i \in G\colon\ P(MS \mid G_i, E_T) = x_i > 0$$

The conditional probabilities: $\{x_i = P(MS \mid G_i, E_T)\}$ and: $\{x = P(MS \mid G, E_T)\}$, are referred to as the "*penetrance*" of MS, during (E_T), respectively, for i^{th} susceptible individual and for the (G) subset. Clearly, the *penetrance* of MS, both for the individual and for the group, will vary depending upon the likelihood of different environmental conditions during different *Time Periods*. If the environmental conditions during some *Time Periods* were such that certain members of the (G) subset have no possibility of ever developing MS then, for these individuals, during these *Time Periods*:

$$P(MS \mid G_i, E_T) = 0$$

Because, currently, some individuals do develop MS, it must be that, during our "current" *Time Period*:

$$P(MS \mid G, E_T) > 0$$

However, if, at some other time, the environmental conditions were such that no member of (G) could ever develop MS then, during these *Time Periods*:

$$P\big(MS \mid G, E_T\big) = P\big(MS \mid E_T\big) = 0$$

We also define a subset (G_{w^s}), which consists of all *women* within the (G) subset {i.e., $(G_{w^s}) = (F, G)$}, and we define the proportion of *women* in (G) as: $p = P(F \mid G)$. In this circumstance, each of the $(m * p)$ *women* in the (G_{w^s}) subset $(d = 1,2, \ldots, mp)$ has a unique genotype (G_{dw^s}). The occurrence of (G_{dw^s}) represents the event that an individual, randomly selected from (Z), belongs to the (G_{dw^s}) subset – a subset consisting of only the dth susceptible *woman* – and the term $\{P(G_{dw^s}) = 1/N\}$ represents the probability of this event. Also, $\{P(G_{w^s}) = mp/N\}$ represents the probability of the event that an individual, randomly selected from (Z), belongs to the (G_{w^s}) subset.

Individuals, who do not belong to the (G) subset, belong to the mutually exclusive (complimentary) subset (G^c), which consists of <u>*all*</u> individuals who have <u>*no*</u> chance, whatsoever, of developing MS, regardless of any environmental experiences that they either have had or could have had. The occurrence of (G^c) is defined as the event that an individual, randomly selected from the population (Z), is a member of the (G^c) subset. The term $P(G^c \mid E_T)$ represents the probability of this event, given the environmental conditions of (E_T). Consequently, each of the $(m^c = N - m)$ "non-susceptible"

individuals in the (G^c) subset $(j = 1,2, \dots, m^c)$ has a unique genotype (G_j). As *above*, the occurrence of (G_j) represents the event that an individual, randomly selected from (Z), belongs to the (G_j) subset – a subset also consisting of a single individual. The term $\{P(G_j) = 1/N\}$ represents the probability of this event. Thus, under <u>*any*</u> environmental conditions, during <u>*any*</u> *Time Period*:

$$\forall G_j \in G^c \ \& \ \forall\ (E_T):\ P(MS \mid G_j, E_T) = P(MS \mid G_j) = 0$$

and thus, $$\forall\ (E_T):\ P(MS \mid G^c, E_T) = P(MS \mid G^c) = 0$$

Notably, "identical" twins, despite having nearly "identical" genotypes (IG), nevertheless, still have subtle genetic differences from each other. Thus, even if these subtle differences are irrelevant to MS susceptibility (as seems likely, and which we assume to be true), these differences still exist. Consequently, every individual – i.e., each complete genotype (G_i) and (G_j) – in the population is unique. Despite this uniqueness, however, we can also define a so-called "susceptibility-genotype" for the ith susceptible individual such that this genotype consists of all (and only) those genetic factors, which are related to MS susceptibility. Because the specification of such a susceptibility-genotype necessarily includes many fewer genetic factors than the ith individual's complete genotype, it is possible that one or more other individuals in the population share the same susceptibility-genotype with the ith individual. For example, in this conceptualization, *MZ*-twins would necessarily belong to the same susceptibility-genotype. We refer to the group of individuals, who belong to the ith susceptibility-genotype, either as the (G_{is}) subset within (Z) or as the "*i-type*" group. Similarly, we refer to individuals who are members of this subset, as "*i-type*" individuals. In this case, the occurrence of (G_{is}) represents the event that a person, randomly selected from (Z), belongs to the (G_{is}) subset, which consists of only a single *i-type*. The term $P(G_{is})$ represents probability of this event. Clearly, because some members of (G) are *MZ*-twins, the total number of such *i-type* groups in the population (m_{it}) will be less than (m). In addition, we define the family $\{G_s\}$ to include all the (G_{is}) subsets (or *i-type* groups) within (Z) and define the event $\{G_s\}$ as representing the union of the disjoint (G_{is}) events such that:

$$\{G_s\} = (G_{1s}) \cup (G_{2s}) \cup \dots \cup (G_{m_{it}s})$$

Because every susceptible person belongs to one and only one of these *i-type* groups, the probability of this event is expressed as:

$$P(\{G_s\}) = P(G) = m/N$$

We also define the set (X) to be the set of penetrance values for members of the (G) subset during some *Time Period*. 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 two mutually-exclusive subsets, $(G1)$ and $(G2)$, suitably defined, such that the penetrance of MS for the subset $(G1)$ during a certain *Time Period* is greater than

that for $(G2)$. The terms $P(G1)$ and $P(G2)$, represent the probabilities of the events that an individual, randomly selected from (Z), is a member of the subsets $(G1)$ and $(G2)$, respectively. Although many such partitions are possible, for the purposes of the present manuscript, $(G1)$ is generally considered interchangeable with the subset of susceptible *women* – i.e., $(G1) = (F \cap G) = (F, G)$ – and $(G2)$ is generally considered interchangeable with the subset of susceptible *men* – i.e., $(G2) = (M \cap G) = (M, G)$.

When considering the enrichment of more penetrant genotypes (*see Methods #2C*), the subsets (F, G) and (M, G) will each be further partitioned into high- and low-penetrance sub-subsets – i.e., $(G1')$ and $(G2')$, respectively – where the basis for this further partition into $(G1')$ and $(G2')$ sub-subsets is unspecified. The definitions of, and the probabilities for, these events mirrors that *above* for $(G1)$ and $(G2)$. Moreover, although the basis for this further partition must be something other than gender, it can be anything else that creates a partition, and it doesn't need to be the same basis for both genders.

{NB: A note on terminology. When a claim refers to any partition of the (G) *subset, the probabilities of developing MS (i.e., the penetrance of MS) for members of the* $(G1)$ *and* $(G2)$ *subsets are designated, respectively, such that:* $x_1 = P(MS \mid G1)$ *and:* $x_2 = P(MS \mid G2)$. *When the partition is based specifically on gender, and to avoid any confusion with our Time Period designations, the group of females/women are indicated, alternatively, either by an upper-case (F) or by a lower-case (w). Similarly, in these circumstances, males/men are indicated, alternatively, either by an upper-case (M) or by a lower-case (m) – see Table 2. In some circumstances (when the meaning is clear), for purposes of notational simplicity, the designations of* (x_1) *and* (x_2) *continue to be used to designate the penetrance of MS for the subsets of susceptible women* (x_1) *and men* (x_2). *In other circumstances, however, greater clarity is provided by using the letter designations. For example, considering the partition of* (G) *into the subsets* (F, G) *and* (M, G)*, the penetrance of MS for susceptible women and men are designated, respectively, as* (Zw) *and* (Zm) *such that:* $x_1 = Zw = P(MS \mid F, G)$ *and:* $x_2 = Zm = P(MS \mid M, G)$ *– see Table 2.*

Moreover, although this manuscript focuses on the gender partition for the disease MS, the Models developed pertain to any partition for any disease, which has data analogous to that found in Canada for the gender partition of MS [6,7].}

B. <u>General Model Specifications and Definitions for Environmental Susceptibility to MS</u>

The term $\{E_i\}$ represents the family of specific sets of environmental exposures, each of which, by itself, is sufficient to cause MS to develop in the ith susceptible individual. Each set within the $\{E_i\}$ family must be distinct (in some respect) from every other set within this family but, otherwise, there can

be any degree of overlap between the factors or events that comprise these sets. Also, there can be any number of sets within the $\{E_i\}$ family although, because $(\forall\, G_i \in G: x_i > 0)$ under <u>*some*</u> environmental conditions, the family cannot be empty. If we assign (v_i) to the number of sets of sufficient exposures for the ith susceptible individual, then $\{E_i\}$ represents the family of sets: $\{E_{i1}, E_{i2}, \dots, E_{iv_i}\}$; and $P(\{E_i\} \mid E_T)$ represents the probability of the event that, at least, one of these sets of sufficient exposures occurs, given the prevailing environmental conditions of the time (E_T). Moreover, if more than one individual belongs to a particular *i-type* group, each group-member will share the same $\{E_i\}$ family of sufficient exposures as the ith individual.

Notably, also, the probability, $P(\{E_i\})$, depends entirely upon the actual environmental conditions that prevail during any *Time Period* – i.e., conditions that are fixed for any specific (E_T). Thus:

$$\forall (i = 1, = 2, \dots, m):\ P(\{E_i\} \mid E_T) \ = A_i$$

where (A_i) represents an unknown constant. This constant (A_i) may be different for each $\{E_i\}$ and, also, it may be different during different *Time Periods*. Consequently, during any (E_T), each $\{E_i\}$ represents a *population-wide* exposure – i.e., an exposure that is "*available*" to everyone. However, whether anyone, in particular the ith susceptible individual, experiences that exposure, is a different matter (*see below*).

Also, for MS to develop in the ith susceptible individual, the events $\{E_i\}$ and (G_i) must occur jointly – i.e., the individual (G_i) must experience at least one of the $\{E_i\}$ environments. This joint occurrence is represented by the subset $(\{E_i\}, G_i)$. The occurrence of $(\{E_i\}, G_i)$ represents the event that an individual, randomly selected from (Z), is both in the (G_i) subset (*described above*) and that they experience an environment sufficient to cause MS in them. The probability of this event, given that this person is a member of (G) subset and given that they are living during (E_T), is represented as $P(\{E_i\}, G_i \mid G, E_T)$. If the event (G_i) occurs without $\{E_i\}$, then whatever exposure does occur, it is insufficient, and the ith individual cannot develop MS. However, the relationship between one individual's family of sufficient exposures to that of others may be complex. For example, every individual (or *i-type*) may have a family with sets unique to them or, alternatively, the families for any two or more individuals (not in the same *i-type* group) may overlap to any degree, even to the point where their families are identical. If <u>*every*</u> susceptible individual has an <u>*identical*</u> family of sufficient environmental exposures, then: $\forall (i): P(\{E_i\}, G_i \mid G, E_T) = P(\{E_i\} \mid G, E_T)$. If some individuals can develop MS under <u>*any*</u> environmental condition, then, for these individuals: $P(\{E_i\}, G_i \mid G, E_T) = P(G_i \mid G, E_T)$. And, finally, if there are (s_e) specific sets of environmental exposure $(e = 1,2, \dots, s_e)$ that are sufficient to cause MS in <u>*any*</u> susceptible individual, then, for the family $\{E_e\}$ of these sets of environmental exposure:

$$\forall G_i \in G \ \ \&\ \ \forall (e):\ P(\{E_e\}, G_i \mid G, E_T) = P(\{E_e\} \mid G, E_T)$$

Definition of the Exposure (E)

Although an individual (genotype) may experience more than one set of exposures, which may be part of one or more than one $\{E_i\}$ family, the individual's total environmental experience is still unique to them. Therefore, we will represent the exposure event of interest (E) as the union of the disjoint events, which exhibit the pairing of susceptible individuals with sufficient environments, such that:

$$(E) = (\{E_1\}, G_1) \cup (\{E_2\}, G_2) \cup \ldots \cup (\{E_m\}, G_m)$$

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

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

Because genotype is assumed to be independent of the prevailing environmental conditions (E_T):

$$\forall G_i \in G\colon\ P(G_i \mid G, E_T) = P(G_i \mid G) = \frac{P(G_i)}{P(G)} = \frac{(1/N)}{(m/N)} = 1/m$$

so that: $$P\big(E \mid G, E_T\big) = (1/m) * \sum_{i=1}^{m} P(\{E_i\} \mid G_i, G, E_T)$$

Thus, the term $P\big(E \mid G, E_T\big)$ represents the probability of the event that a member of the (G) subset, selected at random, will experience an environmental exposure sufficient to cause MS, given their unique genotype and given the prevailing environmental conditions of the time (E_T). Furthermore, from the definition of (E), this event can only occur when the event (G) also occurs, so that:

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

Notably, many environmental factors or events, which are part of a set within the $\{E_i\}$ family, may (and likely do) represent a range of environmental experiences. For example, suppose that Vitamin D deficiency is a factor in one such a set of sufficient exposures. Suppose further that, for the ith susceptible individual to develop MS in response to this set of exposures, they need to experience a vitamin D deficiency of some minimum severity, lasting for some minimum amount of time, and occurring during some critical age-window. In this case, the definition for the environmental event of "vitamin D deficiency" would also include deficiencies of the same (or greater) severity, lasting the same (or a longer) amount of time, and occurring during the same (or narrower) age-window.

Importantly, as noted previously, each set of sufficient environmental exposures is unspecified as to: 1) how many events or factors are involved; 2) when, during the life of an individual, these events or factors need to occur; and 3) what these events or factors are. Each of these sets, of whatever they consist, simply needs to be sufficient, by themselves, to cause MS to develop in the ith (or an *i-type*) susceptible individual.

Partitioning the Environmental Exposure

In addition, any set of environmental exposures, for any individual, can be partitioned conceptually into three mutually exclusive subsets, which we term: (E_{pop}, E_{sib}, and E_{twn}). The subset (E_{pop}) includes all those environmental experiences or events equally likely to be shared by the population generally (including siblings and twins). The occurrence of (E_{pop}) represents the event that a specific environmental event or factor, which an individual experiences, is a member of the (E_{pop}) subbset. The subset (E_{sib}) includes all those environmental experiences or events either more or less likely to be shared by siblings (including twins) compared to the general population. The occurrence of (E_{sib}) represents the event that a specific environmental event or factor, which an individual experiences, is a member of the (E_{sib}) subset. Presumably, the (E_{sib}) environmental experiences occur mostly (but not necessarily exclusively) during childhood. The subset (E_{twn}) includes all those environmental experiences or events more or less likely to be shared by *MZ*- and *DZ*-twins compared both to non-twin co-siblings and to the general population. The the occurrence of (E_{twn}) represents the event that a specific environmental event or factor, which an individual experiences, is a member of the (E_{twn}) subbset. Presumably, the (E_{twn}) environmental events occur mostly (but not necessarily exclusively) during the intrauterine and early post-natal periods. Importantly, creating this partition does not imply that any of these experiences are unique to twins or siblings – *everyone* experiences each environmental component. The difference is that twins and siblings are more or less likely to *share* certain experiences.

For example, each of the (v_i) sets of sufficient environmental exposures within the $\{E_i\}$ family can be partitioned into these three distinct events such that the event (E_{ij}) represents the union of these disjoint events. In this circumstance, therefore:

$$\forall(j = 1,2, \dots, v_i) : \; E_{ij} = (E_{twn})_{ij} \cup (E_{sib})_{ij} \cup (E_{pop})_{ij}$$

where the probability of each event (E_{ij}) is defined as the the joint probability of these three independent component events so that:

$$P(E_{ij}) = P\{(E_{twn})_{ij}, (E_{sib})_{ij}, (E_{pop})_{ij}\} = P(E_{twn})_{ij} * P(E_{sib})_{ij} * P(E_{pop})_{ij}$$

and the event $\{E_i\}$ is represented as:

$$\{E_i\} = (E_{i1}) \cup (E_{i2}) \cup \dots \cup (E_{iv_i})$$

The same is true within every $\{E_i\}$ family – i.e., $\forall(i)$: $(i = 1,2, \dots, m)$.

{NB: Many, perhaps most, environmental exposures are "population-wide" in the sense that the risk of these events is shared by everyone. For example, the amount of sunlight reaching the Earth's surface in a particular region can be considered a "population-wide" exposure in the sense that the same amount of sunlight is "available" to everyone in that region. Despite this, however, there may be certain

individuals or certain subgroups within the population who experience less sun-exposure than others (e.g., if they disproportionately use sun-screen, if they disproportionatley avoid the sun, or if they are otherwise disproportionatley protected from sun-exposure). Conversely, there may also be certain individuals or groups who experience more sun-exposurre than others. However, given the fact that a co-twin (or a non-twin co-sibling) experiences such an imbalance, unless their proband twin (or proband sibling) is either more or less likely to to experience a similar imbalance compared to others, then these exposures would still be considered part of the (E_{pop}) *environment. Also, the* (E_{sib}) *environment may include experiences outside the childhood micro-environment if, for example, sharing the same biological mother made the intra-uterine environment more similar for siblings than that for the general population. In addition, if twins disproportionately shared certain childhood or adult experriences more so than other siblings or the general population, then these experiences woud be part of the* (E_{twn}) *environment.*

Although it is unspecified as to what experiences consitiute each subset, nevertheless, these three subsets of environmental exposure (E_{twn}, E_{sib}, *and* E_{pop}) *are envisioned to be mutually exclusive and that, together, they comprise any idividual's unique environmental experience. Thus, as noted above, every individual experiences each of these components of enviornmental exposure, regardless of whether they are twins or non-twin siblings and regardless of whether they are members of the* (G) *subset. For example, even though the* <u>*same*</u> *intrauterine environment is shared by twins, everyone experiences* <u>*some*</u> *intrauterine environment. Similarly, although both twins and non-twin co-siblings experience a* <u>*similar*</u> *childhood environment, everyone experiences* <u>*some*</u> *childhood environment. Nevertheless, in considering these components of environmental exposure as they relate to the sufficent sets as described above – i.e. for* $(i = 1,2, \dots, m)$ *and for* $(j = 1,2, \dots, v_i)$, *we are here focused on the events* (E_{ij}), *for which, during any Time Period, it will be the case that:*

$$P(E_{ij}) = P(E_{twn})_{ij} * P(E_{sib})_{ij} * (E_{pop})_{ij}$$

Notably, during any specific Time Period (E_T), *both* $P(E_{ij})$ *and its component parts are constants.}*

<u>*Impact of the* (E_{sib}) *Environment*</u>

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 suggest, rather, that the important environmental risks (not considering twins) result from exposures that are experienced *population-wide* [9-15]. Thus, these studies, collectively, provide compelling evidence for the absence of any (E_{sib}) environmental impact on MS.

Relationships between (MS), (E), and (G)

From the definitions of environmental and genetic susceptibility (*above*), for the event of (MS) to occur, both the event (G) and the event (E) must also occur. If either of these events does not occur, the event (MS) cannot occur. Therefore, it is clear that:

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

C. Circumstances Relating to Twins and Siblings of Individuals with MS

The terms (MZ), (DZ), and (S) represent, respectively, the subsets of *MZ*-twins, *DZ*-twins, and non-twin sib-ships within (Z). The occurrence of (MZ), (DZ), or (S) represents the event that an individual, selected at random from (Z), belongs, respectively, to each of these subsets and the terms $P(MZ)$, $P(DZ)$, $P(S)$ represent the respective probabilities of these events. For clarity, the randomly selected individual is always referred to as the "proband twin" or the "proband sibling" depending upon the subset to which they belong. The other member (or members) of the twinship or sibship are always referred to as the "co-twin(s)" or "co-sibling(s)".

Circumstances for Twins and Siblings of Selected Probands

Initially, we will consider two events for an *MZ* twin-pair. The first is the event that the randomly selected proband is a member of the (MS, MZ) subset and their co-twin is a member of the (MZ) subset; the second is the event that the proband is a member of the (MZ) subset and their co-twin is a member of the (MS, MZ) subset. Clearly, the probability of these two events is the same. Therefore, to distinguish the circumstances of the proband from those of the co-twin, we will use the term the term (MZ_{MS}) to indicate, specifically, the status of the co-twin. Thus, during the *Time Period* (E_T):

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

where $\{P(MS \mid MZ, E_T)\}$ represents the probability of the event (MS) in the proband twin during (E_T) and $\{P(MZ_{MS} \mid E_T)\}$ represents the same probability for the event (MS) in the co-twin during (E_T). Also, because any two *MZ*-twins have "identical" genotypes, therefore:

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

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

In which case:

$$\forall G_i \in G\colon\ P\left(G_i \,\middle|\, MZ_{MS}, E_T\right) = \frac{P\left(G_i, MZ_{MS} \,\middle|\, E_T\right)}{P\left(MZ_{MS} \,\middle|\, E_T\right)} = \frac{P(MS, G_i \,|\, MZ, E_T)}{P(MS \,|\, MZ, E_T)} = P(G_i \,|\, MS, MZ, E_T)$$

Every proband who has an *MZ* co-twin in the $(MS, MZ) = (MS, MZ, G)$ subset and who, thus, shares an "identical" genotype with their co-twin, will also be a member of the (G) subset. Therefore, summing over all susceptible individuals:

$$\sum_{i=1}^{m} P\left(G_i, MZ_{MS} \,\middle|\, E_T\right) = P(G, MZ_{MS} \,|\, E_T) = P\left(MZ_{MS} \,\middle|\, E_T\right)$$

and:

$$\sum_{i=1}^{m} P\left(G_i \,\middle|\, MZ_{MS}, E_T\right) = P(G \,|\, MZ_{MS}, E_T) = P(G \,|\, MS, MZ, E_T) = 1$$

Similarly, the term $P(MZ_E)$ represents the probability of the event that the co-twin of an *MZ*-proband, randomly selected from (Z), is a member of the (MZ, E) subsets, respectively. Thus

$$P(MZ_E \,|\, E_T) = P(E \,|\, MZ, E_T)$$

In an analogous manner, for *DZ*-twinships, the status of the co-twin is indicated by the subsets and the events of: (DZ_{MS}) and (DZ_E). And for non-twin sibships, the status of the co-sibling is indicated by the subsets and events of: (S_{MS}) and (S_E).

Thus, the two terms, $P(MS \,|\, MZ_{MS}, E_T)$ and $P(MS \,|\, DZ_{MS}, E_T)$ represent the conditional life-time probability of the event that an individual (the proband), randomly selected from (Z), is a member of the either the (MS, MZ) or the (MS, DZ) subset, given the fact that their co-twin also belongs, respectively, to the (MS, MZ) or the (MS, DZ) subset, and given the prevailing environmental conditions of the time (E_T). These probabilities are estimated by the proband-wise concordance rate for either *MZ*- or *DZ*-twins [16]. This rate is estimated based on the number of concordant twin-pairs (C_{TP}) compared to the number of discordant twin-pairs (D_{TP}) and adjusted based upon the degree to which twins are "doubly ascertained". The term "doubly ascertained", in this context, represents the proportion of twin-pairs, for whom both twins were independently identified by the initial ascertainment scheme [16]. If all twin-pairs are "doubly ascertained" by this scheme, and if the sample from (Z), so ascertained, is random, then the formula for calculating the proband-wise concordance rate is:

$$Twin\ concordance\ rate = (2 * C_{TP})/(2 * C_{TP} + D_{TP}).$$

However, if the probability of "double ascertainment" is less than unity, then this formula requires some modification [16]. In the Canadian data [6] the double-ascertainment rate for concordant *MZ*-twins was 48.5% (16/33).

In a similar manner, the term $P(MS \,|\, S_{MS}, E_T)$ represents the conditional life-time probability of the event that an individual (the proband), randomly selected from (Z), is a member of the (MS, S)

subset, given the fact that one or more of their non-twin co-siblings is a member of the (MS, S) subset and given prevailing environmental conditions of (E_T).

D. Adjustments for the Shared Environment of Twins

Lastly, the term $P(MS \mid IG_{MS}, E_T)$, represents the proband-wise concordance rate for *MZ*-twins during (E_T) – i.e., $P(MS \mid MZ_{MS}, E_T)$ – which has been adjusted for the fact that concordant *MZ*-twins, in addition to sharing their "identical" genotypes (*IG*), also disproportionately share their (E_{twn}) and (E_{sib}) environments with each other. Such an adjustment may be necessary because, if these disproportionately shared environmental experiences contribute to causing MS in the co-twin, they could also increase the likelihood of MS developing in the proband twin and such a circumstance could, potentially, alter our conclusions regarding the nature of genetic susceptibility in the population.

Because, by definition (*see above*):

$$P(MS) = P(MS, E, G) = P(MS, E) = P(MZ_{MS}, E) = P(MZ_{MS}, G) = P(MZ_{MS})$$

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

where: $\forall(i)$: $(i = 1,2, \dots, m)$:

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

In this way, the probability of the proband being a member of the (MS, E, G_i) subset, given the fact that their co-twin is a member of the (MS, MZ) subset, is expressed as: $P(MS, E, G_i \mid MZ_{MS})$. This probability can then be re-expressed as the product of three component probabilities – the probability of MS developing in an *MZ*-proband (G_i) who experiences a sufficient environment, the probability that this *MZ*-proband experiences an environment sufficient to cause MS, and the probability of this *MZ*-proband being a member of the (G_i) subset – where each probability is conditioned on the proband having an *MZ* co-twin, who is a member of the (MS, MZ) subset within (Z).

For susceptible probands who are members of (G), but who may or may not be from twinships or sibships where the co-twin or a co-sibling is a member of the (MS) subset, the analogous probabilities can be written:

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

where: $\forall(i)$: $(i = 1,2, \dots, m)$:

$$P(MS, E, G_i \mid G) = P(MS \mid E, G_i) * P(E \mid G_i) * P(G_i \mid G)$$

Therefore, to estimate the desired "adjusted" probability, $P(MS \mid IG_{MS})$, we need to remove the impact of the shared environment of *MZ*-twins, while leaving the genetic impact of being *MZ*-twins unchanged. Thus, we define:

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

where: $P(E \mid G_i, IG_{MS}) = P(E \mid G_i)$

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

Moreover, we note that the conditioning events (E, G_i) and (E, G_i, MZ_{MS}) both represent the same underlying event for the proband – i.e., the event that the ith susceptible individual (the *MZ*-proband) experiences an environment sufficient to cause MS in them. In this circumstance, therefore:

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

Incorporating this equivalence, into the *above* definitions, yields:

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

or: $P(MS \mid G_i, IG_{MS}) = P(MS \mid E, G_i) * P(E \mid G_i) = P(MS \mid G_i)$

so that: $P(MS, G_i \mid IG_{MS}) = P(MS \mid G_i) * P(G_i \mid MZ_{MS}) = (x_i) * P(G_i \mid MZ_{MS})$

Therefore, summarizing the *above* arguments, the term (IG_{MS}) is defined such that:

$\forall(i)$: $(i = 1,2, \dots, m)$:

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

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

And, therefore, the desired "adjusted" probability, $P(MS \mid IG_{MS})$, can be expressed such that:

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

Effectively, this adjustment equates to a thought-experiment, in which susceptible *MZ*-twins are separated at conception, and in which the proband twin experiences random (E_{twn}), (E_{sib}), and (E_{pop}) environments, just as would any other member of the (G) subset, given the environmental conditions of the time (E_T).

In addition, as discussed previously (*Methods #1B*), the available evidence suggests that the (E_{sib}) environment has no impact on the likelihood of a susceptible individual subsequently developing MS [9-15]. If this evidence is correct, it will also be the case that:

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

where: $P(E \mid G, IG_{MS}) = P(E \mid G, S_{MS}) = P(E \mid G)$

and where: $P(MS \mid E, G, IG_{MS}) = P(MS \mid E, G)$

which represents the probability that the event (MS) occurs for an individual, randomly selected from (Z), given that they are both a member of the (G) subset and that they have an environmental experience sufficient to cause MS in them.

{NB: This definition represents the intended meaning of the "adjusted" proband-wise recurrence rate for MZ-twins – i.e., $P(MS \mid IG_{MS})$*. The next section (which can be skipped) is devoted to estimating an adjustment that matches this definition and that is derived from directly observed epidemiological data.}*

<u>*Estimating the Necessary Adjustment from Observed Data*</u>

Here the event of interest is the development of MS in a randomly selected proband *MZ*-twin, whose co-twin already has or will develop MS – i.e., the co-twin of this proband is a member of the (MZ_{MS}) subset. Consequently, in this conceptualization, the event (IG_{MS}), which is independent of whatever happens to the proband, is envisioned to be identical to the event (MZ_{MS}), defined *above*. Therefore, during any specific *Time Period*, (E_T):

$$P\left(IG_{MS} \mid E_T\right) = P\left(MZ_{MS} \mid E_T\right)$$

The difference, as discussed *above*, is that the probability of the event – i.e., $P(MS \mid IG_{MS})$ – may be different from; $P(MS \mid MZ_{MS})$ due to the shared environmental experiences of *MZ*-twins.

Including the circumstances of both twins and siblings, the occurrences of: (MZ_E), (DZ_E), and (S_E) represent, respectively, the events that a co-*MZ*-twin, a co-*DZ*-twin, or a non-twin co-sibling, experience the exposure (E). The respective probabilities that the probands of these co-twins or co-siblings are members of the (G) subset and that they experience $\{E_i\}$ are expressed as: $P(\{E_i\} \mid MZ_E)$; $P(\{E_i\} \mid G, DZ_E)$; and: $P(\{E_i\} \mid G, S_E)$. Moreover, because any two siblings and any set of *DZ*-twins have the same genetic relationship to each other, and because, by definition, they both disproportionately share the same (E_{sib}) environment, the only difference between them is that *DZ*-twins share the same (E_{twn}) environment whereas siblings don't.

These probabilities are then contrasted to the probability of the event $\{E_i\}$ for a randomly selected member of the (G) subset – i.e., $P(\{E_i\} \mid G)$. As *above*, we can partition each of these exposure events into three independent component events $(E_{twn}, E_{sib},\ \&\ E_{pop})$, during any *Time Period*, such that:

$\forall(i)$: $(i = 1,2, \ldots, m)$:

$$P(\{E_i\} \mid MZ_E) = P(\{E_{twn}\}_i \mid MZ_E) * P(\{E_{sib}\}_i \mid G, S_E) * P(\{E_{pop}\}_i \mid G)$$

$$P(\{E_i\} \mid G, DZ_E) = P(\{E_{twn}\}_i \mid G, DZ_E) * P(\{E_{sib}\}_i \mid G, S_E) * P(\{E_{pop}\}_i \mid G)$$

$$P(\{E_i\} \mid G, S_E) = P(\{E_{twn}\}_i \mid G) * P(\{E_{sib}\}_i \mid G, S_E) * P(\{E_{pop}\}_i \mid G)$$

and: $$P(\{E_i\} \mid G) = P(\{E_{twn}\}_i \mid G) * P(\{E_{sib}\}_i \mid G) * P(\{E_{pop}\}_i \mid G)$$

Probands both from an *MZ*-twinship and those from a *DZ*-twinship experience the same (E_{twn}) environment as their co-twin who either has or will develop MS. Therefore, we assume that:

$$P(\{E_{twn}\}_i \mid MZ_E) = P(\{E_{twn}\}_i \mid G, DZ_E)$$

so that: $$P(\{E_i\} \mid MZ_E) = P(\{E_i\} \mid G, DZ_E)$$

In this circumstance:

$$P(\{E_i\} \mid G, DZ_E)/P(\{E_i\} \mid G, S_E) = P(\{E_{twn}\}_i \mid G, DZ_E)/P(\{E_{twn}\}_i \mid G) = B_i \geq 1$$

so that: $$P(\{E_i\} \mid G, S_E) = P(\{E_i\} \mid G, DZ_E)/B_i$$

where (B_i) is a constant for each (i) during any specific *Time Period* (E_T).

The terms $P(MS, G_i, \{E_i\} \mid G, DZ_E)$ and $P(MS, G_i, \{E_i\} \mid G, S_E)$ represent, respectively, the probabilities of the events that the *i*th individual, randomly selected from (Z), is a member of either the (MS, DZ) or the (MS, S) subsets, given that they are a member of the (G) subset and also given that their co-twin or co-sibling is a member of either the (DZ, E) or the (S, E) subsets. However, by definition (*see Methods #1B*), if both events (G) and $\{E_i\}$ occur together, then the event (MS, G_i) will either occur or not occur, independently of the status of a co-twin or co-sibling.

Therefore, $\forall(i)$: $(i = 1,2, \dots, m)$:

$$P(MS, G_i \mid \{E_i\}, MZ_E) = P(MS, G_i \mid \{E_i\}, G, DZ_E) = P(MS, G_i \mid \{E_i\}, G, S_E) = P(MS, G_i \mid \{E_i\}, G)$$

and, thus: $$\frac{P(MS, G_i, \{E_i\} \mid G, DZ_E)}{P(MS, G_i, \{E_i\} \mid G, S_E)} = \frac{P(MS, G_i \mid \{E_i\}, G, DZ_E)}{P(MS, G_i \mid \{E_i\}, G, S_E)} * \frac{P(\{E_i\} \mid G, DZ_E)}{P(\{E_i\} \mid G, S_E)} = B_i \geq 1$$

In this manner, we can use the adjustment factor (B_i), to remove, mathematically, the impact of the shared $\{E_{twn}\}$ environment for the *i*th individual, during any *Time Period*, such that:

$$P(MS, G_i, \{E_i\} \mid G, S_E) = P(MS, G_i, \{E_i\} \mid G, DZ_E)/B_i$$

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

therefore: $$P(MS, E \mid G, S_E) = \sum_{i=1}^{m} P(MS, G_i, \{E_i\} \mid G, DZ_E)/B_i$$

In this case, we can assign variables (y) and (z) such that:

$$y = 1/B_i \quad \text{and:} \quad z = P(MS, G_i, \{E_i\} \mid G, DZ_E)$$

In which case, the covariance (σ_{yz}), using its standard definition [39], can be expressed such that:

$$\sigma_{yz} = E(yz) - E(y) * E(z)$$

where: $\sum_{i=1}^{m} P(MS, G_i, \{E_i\} \mid G, DZ_E)/B_i = m * E(yz)$

and where: $\sum_{i=1}^{m} P(MS, G_i, \{E_i\} \mid G, DZ_E) = m * E(z) = P(MS, E \mid G, DZ_E)$

Because (*from above*): $\forall(i) : B_i \geq 1$

Therefore: $m * E(yz) \leq m * E(z)$

In this circumstance, we can define an adjustment factor (s_a) such that:

$$E(yz) = E(z)/s_a \leq E(z)$$

where: $s_a \geq 1$

Substituting this back into the definition (*above*) for the covariance (σ_{yz}), yields:

$$\sigma_{yz} = \{(1/s_a) - E(y)\} * E(z)$$

or with rearrangement: $1/s_a = E(y) + \sigma_{yz}/E(z) \leq 1$

Therefore: $P(MS, E \mid G, S_E) = m * E(yz) = m * E(z)/s_a = P(MS, E \mid G, DZ_E)/s_a$

or: $s_a = P(MS \mid G, DZ_E)/P(MS \mid G, S_E) \geq 1$

We define the term (IG_E) in an analogous manner to (MZ_E) – *see above*. Moreover, because everyone who develops MS must experience the event (E), because both *MZ*- and *DZ*-twins disproportionately share the (E_{twn}) and (E_{sib}) environments with their co-twin, and because siblings (including twins) disproportionately share their (E_{sib}) environment with their co-sibling(s), and because everyone shares their (E_{pop}) environment, therefore, with respect to the environmental experiences of the proband, it will be the case that:

$$P(E \mid MZ_E) = P(E \mid MZ_{MS}) = P(E \mid G, DZ_{MS}) = P(E \mid G, DZ_E)$$

Moreover, from the definition of (IG_{MS}) and because the (E_{sib}) environment doesn't seem to contribute to the event (MS, E) and because genotype is independent of (E_T), then, in this circumstance:

$$P(E \mid G, S_E) = P(E \mid G, S_{MS}) = P(E \mid IG_{MS}) = P(E \mid IG_E) = P(E \mid G)$$

Moreover: $P(MS \mid DZ_{MS}, E_T) = P(MS, G \mid DZ_{MS}, E_T) = P(MS \mid G, DZ_{MS}, E_T) * P(G \mid DZ_{MS})$

and: $P(MS \mid S_{MS}, E_T) = P(MS \mid G, S_{MS}, E_T) * P(G \mid S_{MS})$

where: $P(G \mid DZ_{MS}) = P(G \mid S_{MS})$

Therefore, to convert $P(MS \mid MZ_{MS}, E_T)$ into $P(MS \mid IG_{MS}, E_T)$, we can remove the impact of the shared (E_{twn}) environment of *MZ*-twins, during (E_T), using observable population parameters such that:

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

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

{NB: In this development of the adjustment factor (s_a)*, we have assumed that the* (E_{twn}) *environment is the same for MZ- and DZ-twins. We have also assumed, based on evidence [9-15], that the* (E_{sib}) *environment doesn't impact the likelihood of a person subsequently developing MS. However, either one (or both) of these assumptions may be incorrect. For example, certain* (E_{twn}) *exposures may be more or less likely to occur in MZ-twins compared to DZ-twins, in which case our estimate for* (s_a) *would be either an under or an overestimate. Nevertheless, if either the* (E_{twn}) *or the* (E_{sib}) *environments are important to MS pathogenesis, then some adjustment is necessary. It is for this reason that, for our analysis, we consider a wide range of possible values for* (s_a)*, including those circumstances for which no adjustment is necessary – see below.}*

E. <u>*Characterizing Genetic Susceptibility to MS in a Population*</u>

From these *Model* specifications and definitions, we can use estimated values for observable population parameters to deduce the value of the non-observable parameter $P(MS \mid G, E_T)$, which represents the probability of the event that an individual, randomly selected from (Z), will develop MS over the course of their lifetime, given that they are a member of the (G) subset and given the prevailing environmental conditions of (E_T). From the definition of (G), as noted earlier:

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

Therefore, because genotype is assumed to be independent of the prevailing environmental conditions of (E_T):

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

Rearrangement of this equation, yields:

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

Consequently, the value of $P(G)$ can be estimated using the data from <u>*any*</u> specific *Time Period* (E_T) – including ours – during which: $P(MS \mid E_T) > 0$. Therefore, considering only our "current" *Time Period*, this equation can be simplified to yield:

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

Moreover, once the value of $P(G)$ is established, it can then be used to assess the nature of MS pathogenesis. For example, if: $\{P(G) = 1\}$ – i.e., if the penetrance of MS for the population (Z) is the same as that for the subset (G) – then anyone can get MS under the appropriate environmental conditions. By contrast, if: $\{P(G) < 1\}$ – i.e., if the penetrance of MS for the population (Z) is the less than that for the subset (G) – then only certain individuals in the general population (Z) have any possibility of getting MS. In this circumstance, MS must be considered 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, even if MS is "genetic" in this sense, this has no bearing upon whether disease pathogenesis also requires the co-occurrence of specific environmental events.

In this analysis, two basic *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 the epidemiological data obtained during a single *Time Period* (i.e., the "current" *Time Period*). This will be referred to as the *Cross-sectional Model*. The second *Model* takes a longitudinal approach, in which deductions are made both from the "current" epidemiological data and from the observed changes in MS epidemiology that have taken place over the past 4–5 decades [3,6]. This will be referred to as the *Longitudinal Model*. These two *Models* are independent of each other although both incorporate 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 [17-19]. 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* does not make either of these assumptions to estimate possible ranges for the non-observed parameters and several possible conditions for this *Longitudinal Model* are depicted in *Figures 1–4* (*see below*).

For both *Models*, the first step is to assign acceptable ranges for the value of certain "observed" parameters (e.g., twin and sibling concordance rates, the population prevalence of MS, or the proportion of *women* among MS patients). These ranges are assigned such that they always include their calculated 95% *CIs*. However, for certain parameters, the ranges considered plausible are expanded beyond the limits set by the *CIs*. 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), which fit within the acceptable ranges for both the observed and non-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 $\sim 10^{11}$.

2. *Establishing Plausible Ranges for Parameter Values*

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

For notational simplicity, we sometimes use subscripts (*1*) and (*2*) to indicate the parameter values at *Time Period #1* and *Time Period #2* {e.g., $P(MS)_2 = P(MS \mid E_T)$ at *Time Period #2*}. For the purposes of this analysis, those parameter-values observed for persons born between 1976 and 1980 (i.e., *Time Period #2*) are always taken to be the "current" values and, when only this *Time Period* is being considered, the terms (E_T) and the subscript (*2*) are often omitted entirely to simplify the notation.

{NB: In general, for individuals born during Time Period #2 (1976–1980), their MS status cannot be determined until 25-35 years later (i.e., 2001–2015). The estimates of other epidemiological parameters are from reports in the Time Period of (2001–2015), which is also when the Time Period #2 (F:M) sex ratio is assessed [6,7,11-15]. For this reason, Time Period #2 is fixed as the "current" period. However, because the (F:M) sex ratio has increased between every previous 5-year epoch and Time Period #2 [6], the choice of any specific Time Period #1 is equivalent to any other. For our Time Period #1, we chose the 5-year epoch (1941–1945) because it was the earliest epoch with the narrowest CI [6].}

The 2010 Canadian census [20], reported that the proportion of *women* among the general Canadian population (Z) is 50.4%. Thus, *men* and *women* comprise essentially equal proportions of this population and, therefore, the probabilities of the events that an individual, randomly selected from (Z), is a *woman* or a *man* – $P(F)$ and $P(M)$ respectively – are each ~50%. Therefore, by definition:

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

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

Currently, the proband-wise concordance rate [16] for MS in *MZ*-twins, observed in Canada, is:

$$P(MS \mid MZ_{MS})_2 = 0.253$$

with $(n = 146)$ twin-pairs included in this estimation [7]. Therefore, the 95% *CI* for this parameter, calculated from an exact binomial test [21], is:

$$CI = (0.18 - 0.33)$$

The estimates, from different studies, for the "current" proportion of *women* among the MS patients – i.e., $P(F \mid MS)_2$ – in North America ranges between 66% and 76% [3]. For the *Cross-sectional Model*, we expanded the "plausible" range beyond the 95% *CI* calculated from "current" Canadian data presented *below* [6]. Thus, for this *Model*, we considered the range:

$$0.66 \leq P(F \mid MS)_2 \leq 0.78$$

The reason for this is because the "current" estimated range from the Canadian study is quite narrow and some solutions, which fall within the range of different estimates from other locations in

North America [3], might be excluded. This choice permits a wider range of possibilities to be considered as solutions for our *Cross-sectional Model*.

By contrast, for our *Longitudinal Model*, because we were interested specifically in how the parameter $P(F \mid MS)$ has changed for the Canadian population over time [6], we used the 95% *CIs* (from this single study) to estimate the ranges for this parameter value during each *Time Period*. For example, the proportion of *women* among the MS patients in Canada was 69% for patients born during *Time Period #1* (1941–1945) and this proportion was significantly less ($p < 10^{-6}$) than the 76% observed for patients born during the "current" *Time Period #2* (1976–1980) [6]. Although the authors of this study, do not report the actual numbers of individuals in each 5-year epoch, they do report that their 5-year samples averaged 2,400 individuals per epoch [6]. Also, the authors graphically present the 95% *CIs* for the (*F:M*) *sex ratio* during each of these 5-year epochs in the manuscript *Figure* [6]. Estimating that the number of individuals in both *Time Periods #1* and *#2* is ~2,000, and using an exact binomial test [21], for our *Longitudinal Model*, we estimate that:

$$P(F \mid MS)_1 : \quad CI = (0.67 - 0.71)$$

$$P(F \mid MS)_2 : \quad CI = (0.74 - 0.78)$$

Both of these ranges exceed those (based on the 95% *CIs*) presented in the manuscript *Figure* [6].

The "current" proband-wise concordance rates for MS in *female* and *male MZ*-twins, observed in Canada, are 34% and 6.5%, with the total number of *female* and *male* twin-pairs included in these calculations being ($n_1 = 100$) and ($n_2 = 46$), respectively [7]. Using an exact binomial test [21], and using the definitions provided in *Table 2*, the *CIs* for these observations are:

$$y_1' = P(MS \mid F, MZ_{MS})_2: \qquad CI = (0.24 - 0.44)$$

$$y_2' = P(MS \mid M, MZ_{MS})_2: \qquad CI = (0.014 - 0.18)$$

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

$$CI = (y_1' - y_2') \pm 1.96\sqrt{\frac{y_1'(1-y_1')}{n_1} + \frac{y_2'(1-y_2')}{n_2}}$$

In which case: $CI = (0.16 - 0.39)$

Previously, we used three independent methods (*based on observation*) to estimate the value of $P(MS)_2$ [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-of-death study from British Columbia, which reported the proportion of death certificates that mention MS. The parameter-value range supported, collectively, by these different methods was:

$0.0025 \leq P(MS)_2 \leq 0.0046$ [3]. Nevertheless, for the purposes of the present analysis, we expanded the "plausible" range for this parameter to include:

$$0.001 \leq P(MS)_2 \leq 0.006$$

B. Initial Constraints on the Non-Observed Parameter Values

In addition to the observed parameter values (*above*), and using the definitions in (*Table 2*), we determined acceptable values for 12 additional parameters: $P(G)$; $p = P(F \mid \mathrm{G})$; $x = P(MS \mid G)_2$; $x' = P(MS \mid IG_{MS})_2$; $x_1 = P(MS \mid F, G)_2$; $x_1' = P(MS \mid F, IG_{MS})_2$; $x_2 = P(MS \mid M, G)_2$; $x_2' = P(MS \mid M, IG_{MS})_2$ and the ratios: $s_a = P(MS \mid DZ_{MS})_2 / P(MS \mid S_{MS})_2$; $C = P(MS)_1 / P(MS)_2$; $r = x_1'/x_1$; and: $s = x_2'/x_2$

Most of these parameters vary depending upon the level of exposure – i.e., all except $P(G)$ and $P(F \mid \mathrm{G})$. Therefore, the acceptable ranges were estimated for the "current" *Time Period #2*. In several cases, there are constraints on the values that these non-observed parameters can take. For example, $P(MS)$ has been observed to be increasing, especially (but not only) among *women*, in many parts of the world between the two *Time Periods* [6,22-30]. Therefore, the parameter (C) is constrained in three ways such that:

1. $C = P(MS)_1 / P(MS)_2 < 1$

In this case, on theoretical grounds [3], (C) is also constrained such that:

$$C < P(M \mid MS)_2 / P(M \mid MS)_1$$

$$\text{and: } C < P(F \mid MS)_2 / P(F \mid MS)_1$$

Using the limits, provided earlier, for the proportion of *women* among MS patients during different *Time Periods*, the ratio (C) is at its maximum possible value when:

$$P(M \mid MS)_1 = 0.29 \text{ and } P(M \mid MS)_2 = 0.26$$

and, therefore, on these theoretical grounds (*see above*), the value of (C) is further constrained such that:

2. $C < P(M \mid MS)_2 / P(M \mid MS)_1 = 0.26/0.29 = 0.90$

and: 3. $C < P(F \mid MS)_2 / P(F \mid MS)_1 = 0.74/0.71 = 1.04$

Only the second of these constraints satisfies all three, so that the maximum upper bound for (C) is 0.90. Nevertheless, the actual upper bound for (C) will depend upon the values that $P(M \mid MS)_1$ and $P(M \mid MS)_2$ take for in any specific solution. 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 to be an implausibly large increase based on the available data [6,22-30]. Therefore, we conclude that:

$$0.25 \leq C < 0.9$$

Because MS develops in some individuals, the parameter $P(G)$ cannot be equal to 0. Also, because both *women* and *men* can develop MS, the parameter $P(F \mid G)$ cannot be equal to either 0 or 1. Therefore, the plausible ranges for these parameters are:

$$0 < P(G) \leq 1 \text{ and: } 0 < P(F \mid G) < 1$$

Furthermore, the penetrance of MS for the subsets (F, G) and (M, G) can be expressed (*see above*) such that:

$$x_1 = Zw = P(MS \mid F, G) = P(MS, E \mid F, G)$$

$$x_2 = Zm = P(MS \mid M, G) = P(MS, E \mid M, G)$$

Because everyone who develops MS must be a member of the (G) subset, therefore, considering only this subset, the ratio of *women* to *men*, during any *Time Period* can be expressed as:

$$(F{:}M)\ sex\ ratio\} = P(F, MS)/P(M, MS) = P(F, MS, G)/P(M, MS, G)$$

$$\text{or: } (F{:}M)\ sex\ ratio\} = \{P(MS \mid F, G)/(MS \mid M, G)\} * \{P(F \mid G)/P(M \mid G)\}$$

Consequently, using the definitions provided *above* and in *Table 2*, the parameters (x_1), (x_2), (x), (p) and (p'), during any *Time Period* for the gender partition, are related such that:

$$(F{:}M)\ sex\ ratio = \left(\frac{x_1}{x_2}\right) * \left\{\frac{p}{1-p}\right\} = \frac{p'}{1-p'} \qquad \textit{(Equation 1a)}$$

$$\text{or: } (F{:}M)\ sex\ ratio = \left(\frac{Zw}{Zm}\right) * \left\{\frac{p}{1-p}\right\} = \frac{p'}{1-p'} \qquad \textit{(Equation 1b)}$$

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

$$p' = px_1/x \qquad \textit{(Equation 1d)}$$

These relationships require no assumptions and (x_1), (x_2), (x), (p) and (p') must <u>always</u> satisfy *Equations 1a–d* during any *Time Period*, regardless of which *Model* is employed in the analysis [3].

C. <u>Enrichment of More Penetrant Genotypes</u>

In addition, it can be demonstrated that more penetrant genotypes are "*enriched*" in the (G, MS), (F, G, MS) and (M, G, MS) subsets compared, respectively, to the subsets of susceptible individuals (G), (F, G) and (M, G). For example, during any *Time Period*, consider the two ratios, (*A*) and (*B*), such that:

$$A = P(MS \mid F, G)/P(MS \mid G) = x_1/x$$

and: $$B = P(MS \mid M, G)/P(MS \mid G) = x_2/x$$

where: $$P(MS \mid G) = P(MS, F \mid G) + P(MS, M \mid G)$$

Therefore, during any *Time Period*, both of the following relationships must hold:

$$1 = P(F \mid G) + P(M \mid G)$$

and: $$1 = A * P(F \mid G) + B * P(M \mid G)$$

Consequently, during any *Time Period*, it is clear that:

$$A = 1 \quad if\ and\ only\ if: \quad B = 1$$

and: $$A > 1 \quad if\ and\ only\ if: \quad B < 1$$ (and *vice versa*)

Moreover: $$P(F \mid MS, G) = P(F, G) * P(MS \mid F, G)/P(MS, G) = A * P(F \mid G)$$

and similarly: $$P(M \mid MS, G) = B * P(M \mid G)$$

Therefore, if $(A < 1 < B)$ – i.e., if *men* are more penetrant than *women* – then:

$$P(F \mid MS) < P(F \mid G) \quad \& \quad P(M \mid MS) > P(M \mid G)$$

Conversely, if $(A > 1 > B)$ – i.e., if *women* are more penetrant than *men* – then:

$$P(F \mid MS) > P(F \mid G) \quad \& \quad P(M \mid MS) < P(M \mid G)$$

In either case, the more penetrant gender can be described as being "*enriched*" in the (MS) subset in comparison to the (G) subset.

In addition, considered separately, and partitioning each of the subsets (F, G) and (F, G) into their high- and low-penetrance sub-subsets, $(G1')$ and $(G2')$, respectively – *see Methods #1A* – an analogous argument demonstrates that an "*enrichment*" of more penetrant genotypes will also occur within each of these two sub-subsets such that:

$$P(F, G1' \mid MS > P(F, G1' \mid G)$$

and: $$P(M, G1' \mid MS) > P(M, G1' \mid G)$$

Penetrance of MS in Susceptible Men and Women

Importantly, the evidence overwhelmingly favors of the arrangement (*considered above*) in which: $(A > 1)$ or, equivalently, in which: $(Zw_2 > Zm_2)$. For example, the (*F:M*) *sex ratio* has increased by a similar amount (i.e., at a similar rate) between every two *Time Periods* in the Canadian data except one [6]. Thus, the observed increase in MS prevalence over time has disproportionately impacted the

prevalence in *women* [6,22-30]. Indeed, from *Equation 1b*, during the "current" *Time Period*, it must be that:

$$(F{:}M)_2\ sex\ ratio = \left\{\frac{p}{1-p}\right\} * \left(\frac{Zw_2}{Zm_2}\right) = \frac{p'_2}{(1-p'_2)}$$

The value of (p) is unknown but fixed, regardless of *Time Period*, and the values of (Zw_2) and (Zm_2) are also unknown but fixed during the current *Time Period*. The current value of (p') – i.e., (p'_2) – by contrast, is both fixed and known (i.e., observed). From *Equations 1c–d*, and also as indicated *above*, the acceptable value-range for the parameter (p'_2) is such that:

$$0.74 \leq p'_2 = \frac{p*Zw_2}{p*Zw_2+(1-p)*Zm_2} \leq 0.78$$

Therefore, considering the implications of *Equation 1b*, the only circumstances in which the (*F:M*) *sex ratio* can increase over time is, also, for the ratio (Zw/Zm) to increase over time. Certainly, the steady increase in the (*F:M*) *sex ratio* observed in Canada over the last several decades would be most easily explained by those circumstances in which $(p < p')$ during every *Time Period* represented by this data [6]. Moreover, there are two considerations that strongly support this view. First, the only circumstances both for which this ratio will increase and for which: $(Zm_2 \geq Zw_2)$, are also those where: $(p \geq p'_2 \geq 0.74)$ and also those where both $(p \geq p')$ and $(Zm > Zw)$ during every previous *Time Period* represented by the Canadian data [7]. In fact, this conclusion would pertain to every observation throughout the history of MS, including a time when the proportion of *men* among MS patients was reported to exceed that of *women* [40], and also including currently, where women account for a large (and increasing) majority of MS patients [3]. Such a conclusion would certainly be counter intuitive.

Second, and notably, other findings from the Canadian study [7] also suggest the possibility that $(Zm_2 \geq Zw_2)$ is very unlikely. Specifically, this study reported that, for *female* and *male* probands, who have a co-twin in the (MZ_{MS}) subset, the observed penetrance of MS for *women* is (5.7)-fold greater than it is for *men.* Thus, this study [7] reported that:

$$0.34 = P(MS \mid F, G, MZ_{MS})_2 \gg P(MS \mid M, G, MZ_{MS})_2 = 0.065$$

This difference in penetrance between genders is highly significant such that:

$$\chi^2 = 10.5 \quad ; \quad p = 0.001$$

As a consequence of this penetrance difference, and considering *MZ*-twins who either are of unknown concordance or are known to be concordant – i.e., who, respectively, are members of either the (MZ_{MS}) subset or the (MS, MZ_{MS}) sub-subset – this study [7] indicates that *women* are being continuously *enriched* such that:

$$P(F) = 0.5 < (F \mid MZ_{MS}) = 0.66 < P(F \mid MS, MZ_{MS}) = 0.92$$

Because any proband *MZ*-twin is genetically "identical" to their (MZ_{MS}) co-twin, the probands of such co-twins are already "*enriched*" for more penetrant genotypes compared to the general population (*see above*). Consequently, the fact that this "*enriched*" group of *women* has a considerably greater penetrance than similarly "*enriched*" *men* (*see below*), strongly suggests that, currently, susceptible *women* are more penetrant than susceptible *men*. Otherwise, for the circumstance of ($Zm_2 \geq Zw_2$) to be true, the *enrichment* of more penetrant genotypes among *women*, when going from $P(F \mid G)$ to $P(F \mid MS)$, must be substantially greater than it is among *men*, when going from $P(M \mid G)$ to $P(M \mid MS)$. Such a circumstance, using the definitions of *Table 2*, requires that the variance of penetrance values within the subset of susceptible *women* considerably exceed that in and susceptible *men* – i.e., this requires that: ($\sigma^2_{X_1} \gg \sigma^2_{X_2}$) – *see Equations 3a–d* – or, equivalently, it requires that: ($r \gg s$). Either would be hard to rationalize.

Therefore, based on both of these considerations, and without making any assertions regarding the circumstances that might pertain during earlier *Time Periods*, we assume that:

$$x_1 = P(MS \mid F, G)_2 > P(MS \mid M, G)_2 = x_2$$

and, thus: $p < p'_2$

{NB: If our assumption that, currently, ($x_1 > x_2$)*, is incorrect, then, as noted above, susceptible men would always have a greater penetrance than susceptible women. Moreover, this would also indicate that men comprise only a small proportion* ($\leq 26\%$) *of all susceptible individuals or, equivalently, that:* $\{P(G \mid M) \leq 0.36)\}$ *and that:* $\{P(G) \leq 0.68)\}$*. Nevertheless, in such a case, conditions similar to those depicted in* $Figure\ 1C$ *(although approaching a sex ratio greater than unity) would be possible. However, using the definitions from the Longitudinal Model (see below and Table 2), and using the substitution analysis (described above), for every solution, where this assumption was violated, we found that:*

$$P(MS \mid F, G)_2 = P(MS \mid M, G)_2 \;\; and\text{:} \;\; p = p'_2$$

$$P(MS \mid F, G)_1 < P(MS \mid M, G)_1 \;\; and\text{:} \;\; p > p'_1$$

$$R^{app} > 1 \;;\; r > 2;\; r/s \geq 2\;;\; P(G) < 0.25\;;\; and\text{:} \;\; \lambda > 0$$

Notably, if the hazards are proportional, any solution for which ($p = p'_2$) *implies that the response curves in both women and men have already reached their maximum values by Time Period #2 (see Methods #4C; Equation 11d). Such a circumstance further implies that no further increase in the (F:M) sex ratio is possible and, thus, as defined below (see Methods #4A), it also must be that:* ($\boldsymbol{c} \ll \boldsymbol{d}$)*. For our current Time Period to have precisely coincided with this point in time seems surprising, especially*

given the fact that the observed sex ratio has been steadily increasing, and at an approximately even rate, up to this point [6].

Also, as it pertains to other aspects of either Model, it is worth noting that, because: $\{P(F) = 0.5\}$, *<u>any</u> circumstance for which:* $(p \neq 0.5)$, *<u>requires</u> that:* $[P(G) < 1]$ *and, therefore, would indicate that MS is a "genetic" disease in the sense discussed above in Methods #1E.}*

D. <u>Other Constraints on the Non-Observed Parameter Values</u>

Based on theoretical considerations (*see Methods #2C*), we concluded that: $(s_a \geq 1)$. Indeed, this relationship is confirmed observationally, where the recurrence risk of MS for a proband with a co-twin who is a member of the (DZ_{MS}) subset, is consistently reported to be greater than the recurrence risk of MS for a proband sibling with a co-sibling, who is a member of the (S_{MS}) subset [7,31-37]. Therefore, from the definitions of (MZ_{MS}) and (IG_{MS}) – *see Methods #1C & 1D* – it must be the case that, during our "current" *Time Period*:

$$x \leq x' \leq P(MS \mid MZ_{MS})_2$$

$$x_1 \leq x_1' \leq P(MS \mid F, MZ_{MS})_2$$

$$x_2 \leq x_2' \leq P(MS \mid M, MZ_{MS})_2$$

Using the previously described constraints (*above*) on $P(MS \mid MZ_{MS})_2$, $P(MS \mid F, MZ_{MS})_2$ and $P(MS \mid M, MZ_{MS})_2$, therefore, the plausible ranges for (x), (x_1) & (x_2) during the *2nd Time Period* are:

$$0.001 \leq x \leq x' \leq 0.33$$

$$0.001 \leq x_1 \leq x_1' \leq 0.44$$

and: $$0.001 \leq x_2 \leq x_2' \leq 0.18$$

We noted earlier that $P(MS \mid G)$ varies depending upon $P(E)$. Also, as noted earlier in *Methods #1D*, the observed *MZ*-twin concordance rates {i.e., $P(MS \mid MZ_{MS}, E_T)$} may need to be converted into adjusted rates {i.e., $P(MS \mid IG_{MS}, E_T)$} because the observed *MZ*-twin concordance rate will reflect, in part, any increased likelihood that an *MZ*-twin proband will develop MS based on their disproportionately sharing the (E_{twn}) and (E_{sib}) environments with a co-twin who has (or will develop) MS. Because any such increase represents an environmental effect – i.e., it is caused by twins disproportionately sharing the (E_{twn}) and (E_{sib}) environments – the maximum probability of developing MS for susceptible individuals under optimal environmental conditions – i.e., $P(MS \mid G, E)$ – must be greater (at least marginally, if not more so) than these observed *MZ*-twin concordance rates [3]. Consequently, using the *Table 2* notations, from the definitions of $(\boldsymbol{c})$ and $(\boldsymbol{d})$ – see *Methods #4A; below* – because, currently,

$(x_1 > x_2)$, and also because, currently, both $P(MS)$ and $P(F \mid MS)$ are increasing, each of the following relationships must hold simultaneously:

$$P(MS \mid M, MZ_{MS})_2 < \boldsymbol{c} = P(MS \mid M, E, G) \leq \boldsymbol{d}$$

and: $$P(MS \mid F, MZ_{MS})_2 < \boldsymbol{d} = P(MS \mid F, E, G) \leq 1$$

Notably, these relationships *include* the possibility that: $\boldsymbol{c} = \boldsymbol{d} \leq 1$

Finally, as discussed previously (*see Methods #1D* and *above*), we estimate the impact of the disproportionately shared (E_{twn}) and (E_{sib}) environments for *MZ*-twins as:

$$s_a = P(MS \mid DZ_{MS})_2 / P(MS \mid S_{MS})_2 \geq 1$$

and: $$P(MS \mid IG_{MS})_2 = P(MS \mid MZ_{MS})_2 / s_a$$

In the Canadian data [7], the life-time probability of developing MS for the proband of a co-*DZ*-twin with MS (5.4%) was found to be greater than that for the proband of a non-twin co-sibling with MS (2.9%). From these observations, the point-estimate for (s_a) becomes:

$$s_a = P(MS \mid DZ_{MS})_2 / P(MS \mid S_{MS})_2 = 0.0\,54/0.029 = 1.86$$

This point estimate is approximately the same for both *men* and *women* [3]. Thus, these observations from Canada suggest that sharing the (E_{twn}) environment with a co-twin who develops MS markedly increases the likelihood of the proband twin developing MS for both *men* and *women* [3]. Nevertheless, it is possible that impact of these disproportionately shared environments may be over- or under-estimated by the Canadian data [7]. In any event, based on theoretical considerations (*see Methods #1D*), if we use the point-estimate from the Canadian data [7] that:

$$P(MS \mid IG_{MS})_2 \leq P(MS \mid MZ_{MS})_2 = 0.253$$

then: $$P(MS \mid IG_{MS})_2 > P(MS \mid S_{MS})_2 = 0.029 = P(MS \mid MZ_{MS})_2 / 8.7$$

Therefore, for the purpose of both *Models*, we considered the plausible range for (s_a) to be:

$$1 \leq s_a \leq 8.6$$

However, because, the point-estimate for (s_a) from the Canadian data [7] is generally greater than that reported in other similar studies [31-37], we also considered, separately, the more restrictive circumstance in which:

$$1 \leq s_a < 1.9$$

3. ***Cross-sectional Model***:

For the purposes of this section, because the *Cross-sectional Model* considers only the "current" *Time Period #2*, the environmental designations relating to the conditions of the time – i.e., both the

designation of (E_T) and the use of the subscript (*2*) – have been eliminated from those parameter definitions that vary with the environmental conditions of the time (*see Methods #1A; above; see also Table 2*). Also, for simplicity of notation, in the development of the *Cross-sectional Model*, we use the definitions from the previous section, which are also provided in *Table 2*.

Assertions:

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

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

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

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

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

D. $0 \leq \sigma_X^2 \leq (x'/2)^2$

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

Definitions and Assumptions: The subsets (G) and (G^c) have already been defined *(see Methods #1A)* and as noted there:

$$\forall\, G_i \in G: \quad x_i = P(MS \mid G_i)$$

Thus, (x_i) represents the MS-penetrance for the ith individual in the (G) subset during <u>any</u> specific *Time Period* 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 #1A*, we defined the set (X) to include the penetrance value for each of the (m) members of the (G) subset. Thus: $X = \{x_i\}$. Letting (x_G) be a random variable representing any of the $\{x_i\}$ elements within the set (X), and because, as noted earlier, all genotypes in a population are unique, it follows that:

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

$$\forall G_i \in G: \; P\big(G_i \mid G\big) = \frac{P(G_i)}{P(G)} = \frac{(1/N)}{(m/N)} = 1/m$$

$$E(x_G) = \sum_{i=1}^{m} (x_i) * (1/m) = P(MS \mid G) = x \qquad \textit{(Equation 2a)}$$

$$E(x_G^2) = \sum_{i=1}^{m} \big(x_i^2\big) * (1/m) = x^2 + \sigma_X^2 \qquad \textit{(Equation 2b)}$$

$$x' = P(MS \mid G, IG_{MS}) = \sum_{i=1}^{m} P(MS, G_i \mid G, IG_{MS}) \qquad \textit{(Equation 2c)}$$

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

$$x_1 > x_2$$

We now specify our two assumptions:

Assumption #1

Because *MZ*-twinning is generally thought to be non-hereditary [17-19], we assume that every person (i.e., genotype) 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, for any *Time Period*:

$$1.\ \forall G_k \in Z\colon\ P(MZ \mid G_k) = P(MZ)$$

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

$$2.\ \forall G_i \in G\colon\ P(MZ \mid G_i) = P(MZ)$$

If, rarely, *MZ*-twinning were familial [19], it would still be the case that:

$$3.\ \forall G_k \in Z\colon\ P\left(MZ \mid G_k\right) \approx P(MZ)$$

Finally, even if some cases of *MZ*-twinning were thought to be familial [19], the exact relationship (*#2, above*) still follows if those genetic factors, which are associated with *MZ*-twining, are independent of those genetic factors associated with MS. Here we assume that either this or #1 (*above*) pertains.

Assumption #2

The penetrance of MS for a proband *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* co-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}) and (E_{sib}) environments together with an *MZ* co-twin is the same as the impact of experiencing the same (E_{twn}) and (E_{sib}) environments alone. Alternatively, it translates to the testable hypothesis that the mere fact of having an *MZ* co-twin does not alter the (E_{twn}) and (E_{sib}) environments in such a way that MS becomes more or less likely in both twins. Thus, we are here assuming that, for any *Time Period*:

$$\forall G_i \in G\colon\ P(MS \mid G_i, MZ) = P(MS \mid G_i)$$

Proof of Assertion A:

From *Assumption #1*, it follows that:

$$\forall G_k \in Z\colon\quad P(G_k, MZ) = P(G_k) * P(MZ \mid G_k) = P(G_k) * P(MZ)$$

and therefore: $$\forall G_k \in Z\colon\ P\left(G_k \mid MZ\right) = P(G_k, MZ)/P(MZ) = P(G_k)$$

Consequently, also: $\forall G_i \in G\colon\ P\big(G_i \,\big|\, MZ\big) = P(G_i)$

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

$$\forall G_i \in G\colon\ P(G_i, MZ_{MS}) = P\big(MS, G_i \,\big|\, MZ\big) = P\big(G_i \,\big|\, MZ\big) * P\big(MS \,\big|\, G_i, MZ\big)$$
$$= P(G_i) * P(MS \,|\, G_i) = P(MS, G_i)$$

and:

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

This last conclusion is also evident from the definition of (MZ_{MS}) – see *Methods #1C*. From these two equivalences, and from the definition of (IG_{MS}) – see *Methods #1D* – we conclude that, for any *Time Period*:

$$\forall G_i \in G\colon\ P\big(G_i \,\big|\, IG_{MS}\big) = P\big(G_i \,\big|\, MZ_{MS}\big) = \frac{P(G_i, MZ_{MS})}{P(MZ_{MS})} = \frac{P(G_i, MS)}{P(MS)} = P\big(G_i \,\big|\, MS\big)$$

Proof of Assertion B:

From the definition of (IG_{MS}) – see *Methods #1D* – and from *Assertion A,* it directly follows that, for any *Time Period*:

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

Proof of Assertion C:

From the definitions of (G) & (IG_{MS}) – *see Methods #1A & 1D* – it directly follows that:

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

Consequently, during any *Time Period*, the probability $P(MS, G_i \,|\, G, IG_{MS})$ can be re-expressed as:

1. $P(MS, G_i \,|\, G, IG_{MS}) = P(G_i \,|\, G, IG_{MS}) * P(MS \,|\, G_i, G, IG_{MS})$
$$= P(G_i \,|\, G, IG_{MS}) * (x_i)$$

From *Assertion A* and from the definitions of (G) & (IG_{MS}) – *see Methods #1A & 1D* – the term $P(G_i \,|\, G, IG_{MS})$ can be re-expressed as:

2. $P(G_i \,|\, G, IG_{MS}) = P(G_i \,|\, G, MS) = P(G_i, G, MS)/P(MS, G)$

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

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

Combining *1 & 2 (above*) yields:

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

However, from *Equation 2c*, it is the case that:

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

where: $\sum_{i=1}^{m} P(MS, G_i \mid G, IG_{MS}) = \sum_{i=1}^{m} \left(x_i^2\right) * (1/m)/x = E(x_G^2)/x$

Therefore, from *Equation 2b*, it follows that:

$$x' = (x^2 + \sigma_X^2)/x = x + \sigma_X^2/x \qquad \text{(Equation 2d)}$$

Equation 2d 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:

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

Defining the variance of penetrance values within the subsets of susceptible *women* and *men* as ($\sigma_{X_1}^2$) and ($\sigma_{X_2}^2$), respectively, the same line of argument also leads to the conclusions that:

$$x_1 = (x_1'/2) \qquad \text{(Equation 2f)}$$

and: $$x_2 = (x_2'/2) \pm \sqrt{(x_2'/2)^2 - \sigma_{X_2}^2} \qquad \text{(Equation 2g)}$$

Proof of Assertion D:

Equation 2e has real solutions only for the range of:

$$0 \leq \sigma_X^2 \leq (x'/2)^2 \qquad \text{(Equation 3a)}$$

and, also, from *Equations 2f–g*:

$$0 \leq \sigma_{X_1}^2 \leq (x_1'/2)^2 \quad \text{and:} \quad 0 \leq \sigma_{X_2}^2 \leq (x_2'/2)^2 \qquad \text{(Equations 3b)}$$

Notably, the maximum variance for any distribution [38] on the closed interval $[a, b]$ is:

$$\sigma^2 = \{(b - a)/2\}^2$$

Therefore, irrespective of any *Assumptions* we have made (*see above*), the variances ranges provided in *Equations 3a–b* are the maximum possible variance ranges for <u>*any*</u> distribution on each of the respective closed intervals $[0, x']$, $[0, x_1']$ and: $[0, x_2']$.

Also, *Equation 2d* can be re-arranged to yield:

$$\sigma_X^2 = x(x' - x) \qquad \text{(Equation 3c)}$$

and, from *Equations 3b*:

$$\sigma_{X_1}^2 = x_1(x_1' - x_1) \text{ and: } \sigma_{X_2}^2 = x_2(x_2' - x_2) \qquad \text{(Equations 3d)}$$

Assertion C Solutions

Equation 2e has two solutions – the so-called <u>*Upper Solution*</u> and the <u>*Lower Solution*</u>, depending upon the value of the (±) sign. The <u>*Upper Solution*</u> represents the gradual transition from a distribution, when $(\sigma_X^2 = 0)$, in which everyone has a penetrance of (x') to a bimodal distribution, when $\{\sigma_X^2 = (x'/2)^2\}$, in which half of the (G) subset has a penetrance of (x') and the other half has a penetrance of zero. Although, under *some* environmental conditions: $(\forall\, x_i \in X\colon\ x_i > 0)$, as noted previously (*see Methods #1A*), there may be certain environmental conditions, in which, for some individuals in the (G) subset:

$$P(MS \mid G_i, E_T) = 0$$

Therefore, the <u>*Upper Solution*</u>, during any particular *Time Period*, is constrained such that:

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

The <u>*Lower Solution*</u> represents the gradual transition from the bimodal distribution described above to increasingly extreme and asymmetric distributions [3]. The <u>*Lower Solution*</u>, however, is further constrained by the requirement of *Equation 2d* that when: $(\sigma_X^2 = 0)$ then: $(x = x')$. Therefore, the <u>*Lower Solution*</u> is constrained such that:

$$0 < x \leq x'/2$$

Moreover, regardless of whether the <u>*Upper*</u> or <u>*Lower*</u> <u>*Solution*</u> 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*, for the gender partition, during any *Time Period*:

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

and:

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

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

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

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 (F, G) and (M, G), conform to the <u>*Upper Solution*</u> [3]. In this circumstance, on theoretical grounds, from *Equations 2e–g & 3a–b*, it must be the case that either:

$$1 \leq x'/x \leq 2$$

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

Using a "substitution" analysis, we wrote a computer program, which incorporated the above acceptable ranges (*see Methods #2*) 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 *Summary Equations* (*below*) and determining those combinations (i.e., solutions) that fit within the acceptable ranges for both the observed and non-observed parameters (*see Methods #2*).

<u>*Summary Equations*</u>

from *Definitions* & *Equations 1a–d ; 2a–e, 3a–b & 4a–b*

$$P(MS \mid G) = x \;\; ; \;\; P(MS \mid F, G) = x_1 = Zw \;\; ; \;\; P(MS \mid M, G) = x_2 = Zm$$

$$P(MS \mid IG_{MS}) = x' \;\; ; \;\; P(MS \mid F, IG_{MS}) = x_1' \;\; ; \;\; P(MS \mid M, IG_{MS}) = x_2'$$

$$P(F \mid G) = p \;\; ; \;\; P(F \mid MS) = p' \;\; ; \;\; r = x_1'/x_1 \;\; ; \;\; s = x_2'/x_2$$

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

$$x_1 = Zw = \frac{x + \sqrt{x^2 - \{1 + (r/s)(1-p)/p)\}\{x^2 - xx'(1-p)/s\}}}{p + (r/s)(1-p)}$$

$$x_2 = Zm = \frac{x - \sqrt{x^2 - \{1 + (s/r)(p/(1-p)\}\{x^2 - xx'\, p/r\}}}{(1-p) + (s/r)p}$$

$$x = px_1 + (1-p)x_2$$

$$p' = px_1/x$$

$$p'/(1-p') = (x_1/x_2) * \{p/(1-p)\}$$

$$x' = x + \sigma_X^2/x$$

4. ***Longitudinal Model***:

A. <u>*General Considerations*</u>

Following standard survival analysis methods [39], we define the cumulative survival $\{S(u)\}$ and failure $\{F(u)\}$ functions where: $F(u) = 1 - S(u)$. These functions are defined separately for *men* $\{S_m(u)\ and\ F_m(u)\}$ and for *women* $\{S_w(u)\ and\ F_w(u)\}$. In addition, we define the hazard-rate functions for developing MS at different exposure-levels (u) in susceptible *men* and *women* {i.e., $h(u)$ and $k(u)$,

respectively}. These hazard-rate functions for *women* and *men* may or may not be proportional to each other but, if they are proportional, then: $k(u) = R * h(u)$, where $(R > 0)$ represents the hazard proportionality factor. Furthermore, as defined previously (*see Methods #1B*), the term $P(E \mid G, E_T)$ represents the probability of the event that a member of the (G) subset, selected at random, will experience an environmental exposure sufficient to cause MS, given their unique genotype and given the prevailing environmental conditions of the time (E_T). We define the exposure (u) as the odds of this event during the *Time Period* (E_T) such that:

$$u = P(E \mid G, E_T)/[1 - P(E \mid G, E_T)]$$

We define $H(a)$ to be the cumulative hazard function (for *men*) at an exposure-level $(u = a)$ such that:

$$H(a) = \int_0^a h(u)du$$

Similarly, we define $K(a)$ to be the cumulative hazard function (for *women*) at an exposure-level $(u = a)$ such that:

$$K(a) = \int_0^a k(u)du$$

For *men*, following the usual definition of the hazard function [39] that:

$$h(u) = f_m(u)/S_m(u)$$

together with the fact that, by definition:

$$f_m(u) = d[F_m(u)]/du = -d[S_m(u)]/du$$

a standard derivation from survival analysis methods [39] demonstrates that, for *men*, because:

$$h(u)du = -d[S_m(u)]/S_m(u)$$

therefore: $$-H(a) = \int_0^a d[S_m(u)]/S_m(u) = \ln[S_m(a)] - \ln[S_m(0)]$$

and: $$-H(0) = \ln[S_m(0)] - \ln[S_m(0)] = 0$$

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: $\{P(E \mid G, E_T) = 0\}$, no member of (G) can develop MS {i.e., $S_m(0) = 1$}. In this circumstance: $\{\ln[S_m(0)] = \ln(1) = 0\}$; and therefore:

$$S_m(a) = e^{-H(a)}$$

Consequently, the cumulative survival function is exponentially related to the integral of the underlying hazard function – i.e., the cumulative hazard function. Even in the unlikely circumstance that the hazard function is discontinuous at some points, the function will still be integrable in all realistic scenarios. In this circumstance, the failure function for susceptible *men* becomes:

$$F_m(a) = 1 - S_m(a) = 1 - e^{-H(a)}$$

{NB: In this circumstance, we are using the cumulative hazard function, $H(a)$, as a measure of exposure for susceptible men, not as a measure of either survival or failure. By contrast, failure, as defined here, is the event that a person develops MS over the course their life-time. The term: $Zm = P(MS, E \mid M, G, E_T)$ represents the probability that, during the Time Period (E_T), this failure event occurs for a randomly selected individual from the (M, G) subset of (Z). Notably, also, the exposure of $H(a)$ is being used in preference to the, perhaps, more intuitive measure of exposure $(u = a)$ provided above. Nevertheless, when $\{P(E \mid G, E_T) = 0\}$; both exposure measures are zero – i.e., $\{a = 0\}$ and $\{H(a) = 0\}$ – also, as: $\{P(E \mid G, E_T) \to 1\}$; both exposure measures become infinite: i.e., $\{a \to \infty\}$ and $\{H(a) \to \infty\}$ – and, finally, both exposure measures increase monotonically with increasing $P(E \mid G, E_T)$. Therefore, the mapping of the $(u = a)$ measure to the $H(a)$ measure is both one-to-one and onto. Consequently, these two measures of exposure are equivalent and the use of either exposure scale is appropriate. Although the relationship of the $H(a)$ scale to $P(E \mid G, E_T)$ is less obvious than it is for the $(u = a)$ scale where: $P(E \mid G, E_T) = a/(a + 1)$, the $H(a)$ scale, nonetheless, has the advantage that the probability of failure is an exponential function of exposure as measured by $H(a)$ and, thus, it is more mathematically tractable.}

<u>*Environmental Exposure Levels during Different Time Periods*</u>

If the environmental exposure level for susceptible *men* during the *1st Time Period* is $\{H(a_1)\}$, and if, as *above*, we define $\{F_m(a) = Zm\}$ as the failure probability for *men* (i.e., the probability of the event that a randomly selected susceptible *man* develops MS), during some *Time Period* (E_T), and $(\boldsymbol{c})$ as the maximum failure probability for susceptible *men* then:

$$F_m(a) = Zm = P(MS \mid M, G, E_T) = P(MS, E \mid M, G, E_T)$$

and: $$\boldsymbol{c} = \lim_{a \to \infty}(Zm) = P(MS \mid M, G, E) \leq 1$$

In this case, the failure probability for susceptible *men* during the *1st Time Period* (Zm_1), can be expressed as:

$$F_m(a_1) = Zm_1 = P(MS, E \mid M, G)_1 = \boldsymbol{c} * [1 - e^{-H(a_1)}] \qquad \textit{(Equation 5a)}$$

If the exposure level for susceptible *men* during the *2nd Time Period* is $\{H(a_2)\}$, then, because (Zm) is currently increasing with time, the difference in exposure for *men* between the *1st and 2nd Time Periods* can be represented as the difference in the environmental exposure level between these two *Time Periods* (q_m) such that:

$$H(a_2) - H(a_1) = q_m > 0$$

In this circumstance, the failure probability for susceptible *men* during the *2nd Time Period* (Zm_2), can be expressed as:

$$F_m(a_2) = Zm_2 = P(MS, E \mid M, G)_2 = \boldsymbol{c} * [1 - e^{-\{H(a_1)+q_m\}}] \qquad \textit{(Equation 5b)}$$

Rearrangement of *Equations 5a & 5b* yields:

$$1 - Zm_1/\boldsymbol{c} = e^{-H(a_1)}$$

and: $$1 - Zm_2/\boldsymbol{c} = e^{-\{H(a_1)+q_m\}} \qquad \textit{(Equation 5c)}$$

And combining (dividing) these two rearranged *Equations* yields:

$$(1 - Zm_1/\boldsymbol{c})/(1 - Zm_2/\boldsymbol{c}) = e^{q_m} \qquad \textit{(Equation 5d)}$$

or: $$q_m = \ln(1 - Zm_1/\boldsymbol{c}) - \ln(1 - Zm_2/\boldsymbol{c}) \qquad \textit{(Equation 5e)}$$

Previously, we assigned the value of these arbitrary units as ($q_m = 1$) in these *Equations* [3], although such an assignation may be inappropriate. Thus, this unit (whatever it is) still depends upon the actual (but unknown) level of environmental change that has taken place between the two chosen *Time Periods*. From *Equations 5d–e*, this level depends upon the value of ($\boldsymbol{c}$), which can range over the interval of: ($1 \geq \boldsymbol{c} > Zm_2$) – *see Methods #2D*. Because (Zm) increases with increasing exposure, the ratio on the *LHS* of *Equation 5d* is always greater than unity and it increases monotonically as ($\boldsymbol{c}$) varies throughout its range. This ratio is at its minimum when: ($\boldsymbol{c} = 1$) and approaches infinity as: ($\boldsymbol{c} \to Zm_2$). Therefore, we can define (q_m^{min}) as the *"minimum"* possible exposure level change for *men* between these two *Time Periods*. In this case, from *Equations 5d & 5e*, this *minimum* exposure level change will occur under circumstances where:

$$\boldsymbol{c} = P(MS \mid M, E, G) = 1$$

so that, from *Equation 5e*: $$q_m^{min} = \ln(1 - Zm_1) - \ln(1 - Zm_2)$$

However, this *minimum* exposure level change (q_m^{min}) may not accurately characterize the actual (but unknown) level of environmental change, which has taken place for *men* between the two *Time Periods*. Therefore, we will refer to (q_m) as the *"actual"* exposure-level change, which may be different from this *minimum* exposure-level change such that:

$$q_m \geq q_m^{min}$$

In an analogous manner, we also define $\{F_w(a) = Zw\}$ as the failure probability for susceptible *women* during any *Time Period* and ($\boldsymbol{d}$) as the ultimate failure probability for susceptible *women* such that:

$$F_w(a) = Zw = P(MS \mid F, G, E_T) = P(MS, E \mid F, G, E_T)$$

and: $$\boldsymbol{d} = \lim_{a \to \infty}(Zw) = P(MS \mid M, F, E) \leq 1$$

In this case, because (Zw) is also currently increasing with time, the *Equations* for the failure probability in susceptible *women* during the *1st& 2nd Time Periods*, $(Zw_1 \text{ and } Zw_2)$, become:

$$F_w(a_1) = Zw_1 = P(MS, E \mid F, G)_1 = \boldsymbol{d} * [1 - e^{-K(a_1)}] \qquad \textit{(Equation 6a)}$$

and: $$F_w(a_2) = Zw_2 = P(MS, E \mid F, G)_2 = \boldsymbol{d} * [1 - e^{-\{K(a_1) + q_w\}}] \qquad \textit{(Equation 6b)}$$

where $\{K(a_1)\}$ represents the exposure level in *women* at the *1st Time Period* and (q_w) represents the *"actual"* level of environmental change for *women*, which has taken place between the two *Time Periods* such that: $K(a_2) - K(a_1) = q_w > 0$

In a manner directly analogous to that presented *above* for the development of *Equation 5e,* it is also the case that:

$$q_w = \ln(1 - Zw_1/\boldsymbol{d}) - \ln(1 - Zw_2/\boldsymbol{d}) \qquad \textit{(Equation 6c)}$$

Thus, as *above* for susceptible *men*, the *"minimum"* value (q_w^{min}) for the exposure level change in susceptible *women* will occur under those circumstances for which: $(\boldsymbol{d} = 1)$, so that:

$$q_w^{min} = \ln(1 - Zw_1) - \ln(1 - Zw_2)$$

Relationship of Failure to True Survival

However, unlike true survival (where everyone dies given sufficient time), the probability of developing MS, either for the subset of susceptible *women* $\{Zw = P(MS, E \mid F, G)\}$ or for the subset of susceptible *men* $\{Zm = P(MS, E \mid M, G)\}$, may not approach 100% as the probability of exposure $\{P(E \mid G, E_T)\}$ approaches unity. Moreover, the limiting value for the probability of developing MS in susceptible *men* $(\boldsymbol{c})$ need not be the same as that in susceptible *women* $(\boldsymbol{d})$. Also, even though the values of the $(\boldsymbol{c})$ and $(\boldsymbol{d})$ parameters are unknown, they are, nonetheless, constants for any disease process, which requires environmental factors as an essential component of disease pathogenesis, and they are independent of whether the hazards are proportional. Finally, the threshold environmental exposure (at which MS becomes possible) must occur at: $P(E \mid G, E_T) = 0$; for one (or both) of these two subsets, provided that this exposure level is possible [3]. If the hazards are proportional, a difference in threshold (λ) can be 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 and $(\lambda_m = 0)$; if the threshold in *men* is greater than the threshold in *women*, (λ) will be negative and $(\lambda_w = 0)$.

Also, in true survival, both the clock and the risk of death begin immediately at time-zero and continue indefinitely into the future, so that the cumulative probability of death always increases with time. By contrast, here, it may be that the prevailing environmental conditions during some *Time Period*

(E_T) are such that: $P(E \mid G, E_T) = 0$; even for quite an extended period (e.g., centuries or millennia). In addition, unlike the cumulative probability of death, here, exposure can vary in any direction with time depending upon the specific environmental conditions during (E_T). Therefore, although the cumulative probability of failure (i.e., developing MS) increases monotonically with increasing exposure, it can increase, decrease, or stay constant with time.

<u>*Relationship of the (F:M) Sex Ratio to Exposure*</u>

Finally, regardless of (λ), and regardless of any proportionality, during any *Time Period*, the failure probability for susceptible *women* (Zw) can be expressed as:

$$Zw = P(MS, E \mid G, F, E_T) = P(E \mid G, F, E_T) * P(MS \mid E, G, F)$$

or: $$Zw = P(E \mid G, F, E_T) * \boldsymbol{d}$$

and, similarly, the failure probability for susceptible *men* (Zm) can be expressed as:

$$Zm = P(MS, E \mid G, M, E_T) = P(E \mid G, M, E_T) * \boldsymbol{c}$$

Dividing the *2nd* of these two *Equations* by the *1st*, during any *Time Period*, yields:

$$Zm/Zw = \{P(E \mid G, M, E_T)/P(E \mid G, F, E_T)\} * \{\boldsymbol{c}/\boldsymbol{d}\} \qquad \textit{(Equation 7)}$$

Consequently, any observed disparity between (Zw) and (Zm), during any *Time Period*, must be due to a difference between *men* and *women* in the likelihood of their experiencing a sufficient environmental exposure, to a difference between $(\boldsymbol{c})$ and $(\boldsymbol{d})$, or to a difference in both. It is not due to any difference between *women* and *men* in their responsiveness to a sufficient environmental input – at least insofar as this difference is due to something other than the limiting values of $(\boldsymbol{c})$ and $(\boldsymbol{d})$. Under conditions of: $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$, therefore, any difference between (Zw) and (Zm) must be due exclusively to a difference in exposure to sufficient environments between the two genders. Also, regardless of whether the hazards are proportional, and because proportion of *women* among susceptible individuals (p) is a constant (*see Table 2*), therefore, for any solution, the ratio (Zw/Zm), during any *Time Period* (E_T), will be proportional to the observed (*F:M*) *sex ratio* during that period (*see Equation 1b*).

<u>*The Response Curves to Increasing Exposure*</u>

Notably, also, because the response curves for both *men* and *women* are exponential, any two points of observation on these curves will define the entire curve (e.g., the values of Zw and Zm during *Time Period #1 and Time Period #2 – see Equations 5a & 5b* and *6a & 6b*). Moreover, if these two curves can be plotted on the same *x-axis* (i.e., if *men* and *women* are responding to the same environmental events), the hazards are always proportional and the values of $(R = q_w/q_m)$ and (λ) are

determined by a combination of the observed values for (Zw_2), (Zm_2), and the (*F:M*) *sex ratio* change and the fixed (but unknown) values of $(\boldsymbol{c})$, $(\boldsymbol{d})$, and (C) – *see Equations 5e and 6c (above); see also Equation 10c & Summary Equations (below).*

B. <u>*Non-proportional Hazard*</u>

If the hazard functions for MS in *men* and *women* are not proportional, it is always possible that the "*actual*' exposure level changes for *men* and *women* are each at their *"minimum"* values – i.e., (q_m^{min}) and (q_w^{min}) – such that: $(\boldsymbol{c} = \boldsymbol{d} = 1)$.

In this circumstance, the various observed and non-observed epidemiological parameter values still limit possible solutions. However, although $(\boldsymbol{c})$ and $(\boldsymbol{d})$ will still be constants, respectively, for *men* and *women*, no information can be learned about them or about their relationship to each other from changes in the (*F:M*) *sex ratio* and $P(MS)$ over time. The observed changes in these parameter values over time could all simply be due to the different environmental circumstances of different times and different places. In this case, also, although *men* and *women* will still each have environmental thresholds, the parameter (λ) – which relates these thresholds to each other – is meaningless, and there is no hazard proportionality factor (R).

Nevertheless, even with non-proportional hazard, the ratio (Zw/Zm), during any *Time Period*, must still be proportional to the observed (*F:M*) *sex ratio* during that *Time Period* (*see Equation 1b*) and, if: $\boldsymbol{c} = \boldsymbol{d} \leq 1$, then any observed disparity between (Zw) and (Zm), must be due entirely to a difference between *women* and *men* in the likelihood of their experiencing a sufficient environmental exposure (E) during that *Time Period* (*see Equation 7; above*).

C. <u>*Proportional Hazard*</u>

By contrast, if the hazards for *women* and *men* are proportional with the proportionality factor (R), the situation is altered. First, because $(R > 0)$, those changes, which take place for the subsets $P(F, MS)$ and $P(M, MS)$ over time, must have the same directionality. Indeed, this circumstance is in accordance with our epidemiological observations where, over the past several decades, the prevalence of MS has been noted to be increasing for both *women* and *men* [6,22-30]. Second, including a possible difference in threshold between the genders, the proportionate hazard *Model* can be represented by those circumstance for which:

$$\forall\, H(a) > \lambda : \quad K(a) = R * \{H(a) - \lambda\} > 0 \qquad \textit{(Equation 8)}$$

In this case, *Equations 6a & 6b*, which represent the failure probability in susceptible *women* given the exposure during the *1st & 2nd Time Periods*, can be re-written as:

$$Zw_1 = \boldsymbol{d} * \left[1 - e^{-K(a_1)}\right] = \boldsymbol{d} * [1 - e^{-R*\{(H(a_1)-\lambda\}}] \qquad (\textit{Equation 9a})$$

and: $$Zw_2 = \boldsymbol{d} * \left[1 - e^{-K(a_2)}\right] = \boldsymbol{d} * [1 - e^{-R*\{H(a_1)+q_m-\lambda\}}] \qquad (\textit{Equation 9b})$$

Rearranging *Equations 5a & 9a* for <u>*any*</u> *Time Period* yields:

$$1 - Zw/\boldsymbol{d} = e^{-K(a)} = e^{-R*\{H(a)-\lambda\}} \qquad (\textit{Equation 10a})$$

and: $$1 - Zm/\boldsymbol{c} = e^{-H(a)} \qquad (\textit{Equation 10b})$$

Division of *Equation 10a* by *10b*, and rearrangement, during any *Time Period*, yields:

$$\lambda = \{\ln[1 - Zw/\boldsymbol{d}] - \ln[1 - Zm/\boldsymbol{c}]\}/R + [(R-1)/R] * H(a) \qquad (\textit{Equation 10c})$$

Consequently, for those circumstances in which $(R = 1)$, *Equation 10c* becomes:

$$\lambda = \ln[1 - Zw/\boldsymbol{d}] - \ln[1 - Zm/\boldsymbol{c}] \qquad (\textit{Equation 10d})$$

For any specific exposure level $\{H(a) > \lambda\}$, the quantities (Zw) and (Zm) are unknown. However, considering any disease for which a proportionate hazard *Model* is appropriate, the parameters $(\boldsymbol{c}, \boldsymbol{d}, R, \& \lambda)$ are fixed (but unknown) constants, so that, from *Equations 10a & 10b,* the values of (Zm) and (Zw) are also fixed at any specific exposure level $\{H(a)\}$.

<u>*Defining an "Apparent" Proportionality Factor*</u>

We can also define a so-called "*apparent*" value of the hazard proportionality factor (R^{app}) such that: $R^{app} = (q_w^{min}/(q_m^{min})$. This "*apparent*" value incorporates, potentially, two fundamentally different processes. First, it may capture the increased level of "sufficient" exposure experienced by one group compared to the other. Indeed, from *Equation 7*, this is the only interpretation possible for circumstances in which: $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$. Second, however, if we admit the possibility that: $(\boldsymbol{c} < \boldsymbol{d} \leq 1)$, then some of (R^{app}) will be accounted for by the difference of $(\boldsymbol{c})$ from unity. For example, using the *Summary Equations* (*below*), and using the proportionate hazard *Model*, (q_m) – the *"actual"* exposure level change in *men* – when $(\boldsymbol{d} = 1)$, has the limits:

$$q_m^{min} \leq q_m \leq q_w^{min}$$

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

From this, we can define the *"actual"* hazard proportionality factor (R), at $(\boldsymbol{d} = 1)$, such that:

$$R^{app} \geq R = q_w^{min} \,/\, q_m$$

In this manner, if $(q_m > q_m^{min})$, a portion of the "*apparent*" value (R^{app}) will be accounted for by a reduction in value of $(\boldsymbol{c})$ from unity, if such a reduction is possible. Moreover, if such a reduction is possible for susceptible *men*, then, clearly, it is also possible that the value of $(\boldsymbol{d})$ is also reduced from

unity in susceptible *women*, in which case: $(\boldsymbol{c} < \boldsymbol{d} < 1)$, where the *"actual"* exposure level in *women* (q_w) would be greater than its minimum value (q_w^{min}) such that:

$$R = q_w/q_m > q_w^{min}/q_m$$

Consequently, in each of these circumstances, the *"actual"* value of (R) may be different from its *"apparent"* value (R^{app}) although, from *Equations 7, 9a–b & 10a–b*, under the conditions of: $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$ then: $(R = R^{app})$.

{NB: Considering Equations 5a & 9a for any Time Period (above), the $(\boldsymbol{c})$ *and* $(\boldsymbol{d})$ *constants, although defined differently (see above), are each the equivalent of a y-axis "scaling factor" for their respective exponential curves, considered separately. However, when* $(\boldsymbol{c} = \boldsymbol{d})$*, this "scaling factor" is the same for both men and women and, therefore, under those conditions where:* $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$*, all response curves for both men and women, which have the same* (R) *and* (λ)*, will differ from each other only in that the y-axis scale is different. For example, suppose that we define a constant* (α) *such that:* $\boldsymbol{c} = \boldsymbol{d} = \alpha < 1$*. In this case, the response curves for men and women, depicted after changing the y-scale by multiplying both sides of Equations 5a & 9a (for any Time Period) by the constant* $(1/\alpha)$*, are identical to those curves having the same* (R) *and* (λ)*, but depicted for:* $(\boldsymbol{c} = \boldsymbol{d} = 1)$.

This does not imply that these response curves are "actually" identical. Rather, the fact that the curves for those conditions under which: $(\boldsymbol{c} = \boldsymbol{d} = \alpha < 1)$ *can be scaled to be identical to those depicted for:* $(\boldsymbol{c} = \boldsymbol{d} = 1)$*, only indicates that the relationship* <u>*between*</u> *the response curve for men and that of women is the same regardless of the value of* (α) *– i.e., any changes in the* (Zw/Zm) *ratio with increasing exposure or, equivalently, any changes of the F:M sex ratio (see below, see also Equation 1b), will be the same for any value of* (α)*. When* $(\boldsymbol{c} < \boldsymbol{d})$*, this kind of transformation is not possible.}*

<u>*Implication that the* (R) *Value has for the Value of* (λ)</u>

We define the ratios $(C_F \ \& \ C_M)$ – *see Table 2* – and note that, because both $P(MS)$ and the (*F:M*) *sex ratio* are both increasing with time [6,22-30], therefore:

$$C_F = P(F,MS)_1/P(F,MS)_2 < P(M,MS)_1/P(M,MS)_2 = C_M$$

Notably, also, from *Equation 1b*, during any *Time Period*:

$$(F{:}M)\ sex\ ratio = (Zw/Zm) * \{p/(1-p)\}$$

where (p) is independent of the environmental conditions of any *Time Period*. Therefore, for all solutions, the ratio (Zw/Zm) will mirror (*F:M*) *sex ratio* (i.e., changes in both ratios will have the same directionality).

For Conditions where: $(R = 1)$:

Under those circumstances for which: $(R = 1)$, it must be the case that:

$$q_m = q_w \geq q_w^{min}$$

When: $(\lambda = 0)$, from *Equation 10d*:

$$Zm/\boldsymbol{c} = Zw/\boldsymbol{d}$$

or: $$Zw/Zm = \boldsymbol{d}/\boldsymbol{c} \qquad \text{(Equation 11a)}$$

Therefore, in this circumstance, the *F:M sex ratio* will remain constant, regardless of the exposure level.

However, using the previously derived *Summary Equations* [3] presented at the end of this section:

$$\boldsymbol{d}/\boldsymbol{c} = \{Zw/Zm\} * \{(e^{q_w} - C_F)/(e^{q_w} - C_M)\} > Zw/Zm$$

This relationship indicates that: $Zm/\boldsymbol{c} > Zw/\boldsymbol{d}$

So that, from *Equation 10d*: $\lambda > 0$

Thus, if $(R = 1)$ and the *F:M sex ratio* is increasing, it must be the case that the threshold in susceptible *women* is greater than that in susceptible *men*.

For Conditions where: $(\lambda \leq 0)$ & $(R > 1)$:

Because $(H(a) \geq 0)$, therefore, from *Equation 10c*, under these conditions, for any *Time Period*:

$$\{\ln(1 - Zw/\boldsymbol{d}) - \ln(1 - Zm/\boldsymbol{c})\}/R = \lambda - [(R-1)/R] * H(a) \leq 0$$

or: $$\ln(1 - Zw/\boldsymbol{d}) - \ln(1 - Zm/\boldsymbol{c}) \leq 0 \qquad \text{(Equation 11b)}$$

In turn, *Equation 11b,* in this circumstance, requires that:

$$Zm/\boldsymbol{c} \leq Zw/\boldsymbol{d}$$

or: $$Zw/Zm \geq \boldsymbol{d}/\boldsymbol{c} \qquad \text{(Equation 11c)}$$

Moreover, regardless of the value of (R), from the definitions of (E), $(\boldsymbol{c})$, and $(\boldsymbol{d})$, and from *Equation 7* (*above*):

$$\lim_{a \to \infty} (Zw/Zm) = \boldsymbol{d}/\boldsymbol{c} \qquad \text{(Equation 11d)}$$

Therefore, because both (Zw) and (Zm) increase monotonically with increasing exposure (*see Methods #4A; above*), and because $(R > 1)$, and because $(H(a) \geq 0)$, the condition that:

$$Zw/Zm \geq \boldsymbol{d}/\boldsymbol{c}$$

requires that: $Zw_1/Zm_1 \geq Zw_2/Zm_2 \geq \boldsymbol{d}/\boldsymbol{c}$:

so that, in this circumstance, the ratio (Zw/Zm) decreases or remains constant with increasing exposure. Because the (Zw/Zm) ratio mirrors the *F:M sex ratio* (*see above*), therefore, under these conditions, the

F:M sex ratio will also be decreasing or remaining constant (e.g., *Figures 1A–B*) – a conclusion, which is contrary to all the available evidence [6,22-30]. Consequently, the conditions of: $(R > 1)$ & $(\lambda \leq 0)$ are not plausible, given the Canadian data [6].

Combining these two conclusions, therefore, it is clear that:

$$\forall(R \geq 1):\ \lambda > 0$$

Thus, based solely on the observations of an increasing $P(MS)$ and an increasing (*F:M*) *sex ratio* with time [6,22-30] – a circumstance which is true considering the "current" *Time Period #2* together with any of the reported previous 5-year epochs as *Time Period #1* [6] – we can conclude, based on purely theoretical grounds, that, if the hazards are proportional, and if: $(R \geq 1)$, susceptible *women* must have a higher threshold than susceptible *men*.

<u>*For Conditions where*</u>*:* $(\lambda \geq 0)$ & $(R \leq 1)$:

If: $(\lambda \geq 0)$ & $(R \leq 1)$ & $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$; then *men* would have as great (or a greater) failure probability than *women* (i.e., $Zm \geq Zw$) at every exposure level (e.g., *Figure 1C*). However, as noted in *Methods #2C* (*above*), currently:

$$Zw_2 = P(MS, E \mid F, G)_2 > P(MS, E \mid M, G)_2 = Zm_2$$

so that the posited conditions are not possible. Nevertheless, under the conditions of $(\lambda \geq 0)$ & $(R \leq 1)$, it is still possible that: $(\boldsymbol{c} < \boldsymbol{d} \leq 1)$ – e.g., *Figure 1D*.

<u>*For Conditions where*</u>*:* $(\lambda < 0)$ & $(R \leq 1)$:

In this circumstance, *Equation 11b* still holds so that; if: $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$, then: $(Zm > Zw)$ so that, again, the posited conditions are not possible although it is still possible that: $(\boldsymbol{c} < \boldsymbol{d} \leq 1)$.

Thus, based on these latter two conditions, it is clear that:

$$\forall(R \leq 1):\ \boldsymbol{c} < \boldsymbol{d} \leq 1$$

D. <u>*Strictly Proportional Hazard*</u>*:* $(\lambda = 0)$

If the hazards in *men* and *women* are "strictly" proportional to each other, then it must be the case that: $(\lambda = 0)$. Therefore, when: $(|\lambda| > 0)$, as it must be when $(R \geq 1)$, the hazards cannot be "strictly" proportional. Indeed, for those circumstances in which $(R \geq 1)$ and $(\lambda = 0)$, the observed (*F:M*) *sex ratio*, as discussed *above*, either decreases or remains constant with increasing exposure (*see Method #4 Equations 11a–c*), regardless of the parameter values for $(\boldsymbol{c})$ and $(\boldsymbol{d})$ – e.g., *Figures 1A & 1B*. Consequently, the only "strictly" proportional circumstances, which are possible, are those in which *men*

have a greater hazard than *women* – i.e., $(R < 1)$. Moreover, if *men* have a greater hazard than *women*, then, as noted *above*, the conditions of: $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$ & $(\lambda = 0)$ are also excluded.

Thus, the only possible "strictly" proportional circumstances are those, in which both $(R < 1)$ and $(\boldsymbol{c} < \boldsymbol{d} \leq 1)$ – e.g., *Figure 1D.*

{NB: In these and subsequent Figures, all response curves exemplifying the conditions in which $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$*, are depicted for the condition* $(\boldsymbol{c} = \boldsymbol{d} = 1)$*. Nevertheless, for all those conditions where* $(\boldsymbol{c} = \boldsymbol{d} < 1)$*, the response curves differ from the curves depicted in the Figures only in so far as the y-axis has a different scale. Therefore, the response curves, depicted at:* $(\boldsymbol{c} = \boldsymbol{d} = 1)$*, are representative of all curves for which* $(\boldsymbol{c} = \boldsymbol{d})$ *– see Methods #3C.}*

E. <u>*Intermediate Proportional Hazard:*</u> $(\lambda < 0)$

We can also consider another possible *Model*, which is intermediate between the "strictly" proportional and non-proportional hazard *Models* discussed *above*. In this intermediate *Model*, the hazards are still held to be proportional although the onset of the response curves are offset from each other by an amount $(\lambda \neq 0)$. As noted earlier: $\forall\ (R \geq 1)$: $\lambda > 0$. Therefore, for those circumstances in which $(\lambda < 0)$, the hazard in *men* must be greater than the hazard in *women*. Otherwise, the (*F:M*) *sex ratio* will decrease with increasing exposure, which is contrary to the evidence [6] – e.g., *Figure 2A*. Moreover, under those conditions, for which $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$ & $(R < 1)$ & $(\lambda < 0)$, the (*F:M*) *sex ratio* will initially decrease with increasing exposure to a level below $\{p/(1-p)\}$ and then, this ratio will steadily increase back to this ratio – e.g., *Figure 2B*. However, such a circumstance requires that $(Zm > Zw)$ throughout the entire response curve until an (*F:M*) *sex ratio* of: $\{p/(1-p)\}$ is reached. This is also contrary to the evidence where currently $(Zm_2 < Zw_2)$ and, thus, where: $\{(F{:}M) > p/(1-p)\}$ – *see Equation 1b; Methods #2C*. Thus, the only circumstances, in which $(\lambda < 0)$ is possible, are those where: $(\boldsymbol{c} < \boldsymbol{d})$ – e.g., *Figure 2D* – but, even then, this is only possible in limited circumstances – e.g., *Figures 2C–D*.

F. <u>*Intermediate Proportional Hazard:*</u> $(\lambda > 0)$

By contrast, when $(\lambda > 0)$, although the hazard is still proportional, there are no absolute constraints on the hazard in *men* relative to that in *women*. Thus, the conditions of: $(R < 1)$ & $(\lambda \geq 0)$ – e.g., *Figure 1D* – and: $(R \geq 1)$ & $(\lambda > 0)$ – e.g., *Figure 3C* – can each lead to very similar conclusions. For this discussion, we define (G_{ia}) to be the susceptibility genotype of the ith susceptible individual that includes all (and only) those genetic factors, related to MS susceptibility, which are located on autosomal chromosomes. Also, the occurrence of (G_{ia}) represents the event that an autosomal genotype, randomly

selected from all such genotypes within (Z), is a member of the (G_{ia}) subset. We define the family $\{G_a\}$ to include all the subsets (G_{ia}) within the (G) subset. In a similar manner, the occurrence of $\{G_a\}$ represents the event that an autosomal genotype, randomly selected from all such genotypes within (Z), is a member of the $\{G_a\}$ family. The "susceptibility" genotype (G_{is}), defined previously (*Methods #1A*), includes all (and only) those genetic factors, which are related to MS susceptibility (located on any chromosome) but does not include the entire genotype of the ith individual. The event (G_{is}) and the family $\{G_s\}$ have also been defined previously (*see Methods #1A*). Naturally, *women* and *men* may be members of the same (G_{ia}) subset, but not be members of the same (G_{is}) subset, either if there are factors on the *X*-chromosome related to susceptibility that differ between some susceptible *men* and *women*, or if there are factors on the *Y*-chromosome in some *men*, which are related to susceptibility.

Considerations of Exposure "Intensity"

To evaluate the implications of circumstances in which $(\lambda > 0)$, we also need to develop a notion of what we will call the "*intensity*" of exposure. For example, suppose that *every* set of sufficient exposures for *every* susceptible individual (of any *i-type*) includes both a deficiency of vitamin D and an Epstein Barr Viral (*EBV*) infection [3], each event needing to occur during some (but not necessarily the same) critical period during a person's life [8]. Moreover, suppose further that, to cause MS, the vitamin D deficiency needs to be more severe or longer lasting in *women*, or that the critical age-window for the *EBV* infection is narrower in *women* than *men*, or both. In each of these circumstances, we can describe susceptible *women* as requiring a greater "*intensity*" of exposure than susceptible *men* and any threshold difference (λ) will reflect this increased "*intensity"* required by susceptible *women*.

This framing, however, does not imply that susceptible *women* require a more "*intense*" exposure to every environmental factor within a sufficient set. Nor does it imply that every observed change in environmental exposure will demonstrate this altered threshold. Thus, for example, suppose that (in the *above* example) susceptible *men* and *women* require the same "*intensity*" of *EBV* exposure but that *women* require a more "*intense*" vitamin D deficiency. In this circumstance, if the environmental change that occurs is due, in whole or in in part, to a change in population vitamin D levels, then the threshold difference will be apparent. By contrast, if the environmental change that occurs only involves only a change in *EBV* exposure, no threshold difference will be identified although, also, no change in the *F:M sex ratio* will be found (*see Methods #4C; Equations 11a-c*). Nevertheless, for those circumstances in which the hazards are proportional and: $(\lambda > 0)$, we can be confident that the environmental change, which took place between the two *Time Periods*, include changes in a factor (or factors) for which *women* require a more *"intense*" exposure.

Also, if *men* require a less "*intense*" exposure to develop MS than *women*, presumably, in such a circumstance, both genders can still develop MS in response to the more "*intense*" exposures required by *women* (*see Methods #1B* for a discussion regarding environmental factors representing a range of exposures). In this case, we also need to expand our definition of the exposure family $\{E_i\}$. Thus, because (G_{ia}) genotypes are purely autosomal, we expect that:

$$P(G_{ia} \mid M) = P(G_{ia} \mid F)$$

$$P(G_{ia}, F, G_{is}) = P(F, G_{is})$$

and: $$P(G_{ia}, M, G_{is}) = P(M, G_{is})$$

However, as noted *above*, these equivalences do not imply either that:

$$P(F, G_{is}) = P(M, G_{is})$$

or that both: $P(F, G_{is}) > 0$ and: $P(M, G_{is}) > 0$

In this conceptualization, the environmental factors that comprise each set of sufficient exposures within the $\{E_i\}$ family (for *i-type* individuals) are the same regardless of whether the *i-type* individual happens to be a *man* or a *woman*. The difference, however, is that *women*, compared to *men*, require a more "*intense*" exposure to one or more of the factors that comprise these sets. We designate this subset of more "*intense*" exposures, within $\{E_i\}$, as $\{E_{iw}\}$. The subset of (E), which includes only the more "*intense*" environmental exposures required by *women* will be designated as (E_w). By this definition, therefore: $(E_w) \subset (E)$; so that

$$P(E \mid G, E_T) \geq P(E_w \mid G, E_T).$$

Following the arguments used in the development of *Equation 7* (*Methods #4A, above*) and, if *women* require a more *"intense"* exposure than *men* {i.e., for $(\lambda > 0)$ *as envisioned above*}, then, during any *Time Period*:

$$Zw = P(MS, E_w \mid G, F, E_T) = P(E_w \mid G, F, E_T) * \boldsymbol{d}$$

and: $$Zm = P(E \mid G, M, E_T) * \boldsymbol{c} \geq P(E_w \mid G, M, E_T) * \boldsymbol{c}$$

And, again, we can divide the *1st* of these two *Equations* by the *2nd*, to yield:

$$Zw/Zm = \{P(E \mid G, F, E_T)/P(E_w \mid G, M, E_T)\} * \{\boldsymbol{d}/\boldsymbol{c}\} \qquad \textit{(Equation 12a)}$$

Consequently, for any circumstance in which: $\boldsymbol{c} = \boldsymbol{d} \leq 1$:

$$Zw/Zm \leq P(E_w \mid G, F, E_T)/P(E_w \mid G, M, E_T) \qquad \textit{(Equation 12b)}$$

Therefore, by assuming that: $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$, we are also assuming that any difference in disease expression between susceptible *men* and *women* is due entirely to a difference between susceptible *men* and *women* in the likelihood of their experiencing a sufficient environmental exposure, despite the fact

that, for every (i), the probability of an exposure to $\{E_i\}$ and $\{E_{iw}\}$ – i.e., $P(\{E_i\} \mid E_T)$ and $P(\{E_{iw}\} \mid E_T)$, respectively – are both *population-wide* and fixed during any specific *Time Period* (E_T). Because such exposures are "*available*" to everyone, therefore, if the level of sufficient exposure differs between genders, any such difference will be due to a difference in *behavior* between susceptible *women* and *men* – i.e., to an increased exposure to, or avoidance of, susceptible environments by one or the other gender (perhaps consciously or unconsciously; or perhaps structurally, such as being due to differing gender-roles, differing occupations, differing recreational activities, etc.).

<u>*Exposure Considerations for i-Type Individuals*</u>

If both *men* and *women* are (or potentially could be) members of any specific *i-type* group, by definition, these *men* and *women* both have a non-zero probability of developing MS in response to every one of the (v_i) sufficient sets of exposures within the $\{E_i\}$ family for this group. As discussed earlier, in these circumstances, these specific *i-types*, considered separately, will necessarily exhibit proportional hazards for the two genders (*see Methods #4A; above*). We have already defined the subset (G_{w^s}) – *see Methods #1A* – which includes all susceptible *women* (G_{dw^s}) in (Z), where $(d = 1,2, \ldots, mp)$. If <u>*every*</u> *i-type* group includes (or, potentially, could include) both *men* and *women*, then, at every exposure level for a *man* $\{H(a) > \lambda\}$, we can define a proportionality constant $(R_i > 0)$, such that the exposure level for any *i-type* susceptible *woman* $\{K_i(a) > 0\}$ can be expressed as:

$$\forall\{G_{dw^s} \in (F, G_{is}):\ \ K_i(a) = R_i * \{H(a) - \lambda\}$$

{NB: Here, we don't need to consider the i-type specific exposure for men, $H_i(a)$*, because, by definition, it is always true that, if every i-type has the same threshold* $(\lambda > 0)$ *then, for all* $\{H(a) > \lambda\}$ *and for all* (i)*, both* $\{H(a) - \lambda > 0]\}$ *and* $\{K_i(a) > 0\}$*. Therefore, in this circumstance, there will be some constant* $(R_i > 0)$ *that permits this statement to be true for each* (i)*. See below for a consideration of the impact of different i-types having different thresholds.}*

Also, because each (G_{dw^s}) is a member of one or another of these posited *i-type* groups, we can define an exposure level $\{K_{dw^s}(a)\}$ and a proportionality factor $\{R_{dw^s}\}$ for each susceptible *woman* such that:

$$\forall\ G_{dw^s} \in (F, G_s):\ \ K_{dw^s}(a) = R_{dw^s} * (H(a) - \lambda)$$

where: $\forall\ G_{dw^s} \in (F, G_{is}):\ K_{dw^s}(a) = K_i(a)\ \ and:\ \ R_{dw^s} = R_i$

Thus, the expected exposure level for susceptible *women* can be expressed such that:

$$K(a) = E\{K_{dw^s}(a)\} = \sum_{d=1}^{mp} R_{dw^s} * \{H(a) - \lambda\}/mp = R * \{H(a) - \lambda\}$$

where: $R = E(R_{dw^s})$

Consequently, if *women* and *men* can, potentially, be members of every *i-type* group, the hazards for *women* and *men* will always be proportional, although the hazard proportionality factor (R_i) need not be the same for every *i-type* group. Alternatively, if *men* and *women* are of different *i-types*, but both are responding to the same environmental events, the hazards, again, will be proportional. Additionally, it is possible that the difference in threshold between *women* and *men* (λ_i) may be different for individuals of different *i-types*. We consider, first, this possibility for those circumstance in which: $(\lambda > 0)$. In this circumstance $\{\lambda = \min(\lambda_i) > 0\}$ because, by definition, some *women* will begin to develop MS at this level of exposure. We can also define the threshold difference (λ_{dw^s}) between each susceptible *woman* and that of susceptible *men*, in which case, because $(\lambda > 0)$, by definition: $(\lambda_m = 0)$ – *see Methods #4A*. Thus:

$$\forall\, G_{dw^s} \in (F, G_{is}):\ \ \lambda_{dw^s} = \lambda_i$$

In this manner, the proportionality constants for each *i-type* $(R_i > 0)$ and each *woman* $(R_{dw^s} > 0)$ can be replaced by a *"adjusted"* proportionality constants $(R'_i > 0)$ and $(R'_{dw^s} > 0)$ such that:

$$\forall\, G_{dw^s} \in (F, G_s):\ \ K_{dw^s}(a) = R_{dw^s} * (H(a) - \lambda_{dw^s}) = R'_{dw^s} * \{H(a) - \lambda\}$$

where: $\forall\, G_{dw^s} \in (F, G_{is}):\ K_{dw^s}(a) = K_i(a)\ ;\ \ R_{dw^s} = R_i\ \ ;\ \ R'_{dw^s} = R'_i$; and: $\lambda_{dw^s} = \lambda_i$

Thus, in this case, the expected exposure level for susceptible *woman* can be expressed such that:

$$K(a) = E\{K_{dw^s}(a)\} = \sum_{d=1}^{mp} R'_{dw^s} * \{H(a) - \lambda\}/mp = R * \{H(a) - \lambda\}$$

where: $R'_{dw^s} = R_{dw^s} * \{(H(a) - \lambda_{dw^s})/(H(a) - \lambda)\} \leq R_{dw^s}$

and where now: $R = E(R'_{dw^s})$

For the circumstance where: $(\lambda < 0)$, the analysis is only changed in that: $\{\lambda = \max(\lambda_i)\}$ and that, in this circumstance: $(\lambda_w = 0)$. Thus, in either case, the hazards will still be proportional.

By contrast, if *men* and *women* are each responding to different environmental events, the hazards will not be proportional. In such a circumstance, *men* with MS would be envisioned as having a disease distinct from MS in *women*. Alternatively, perhaps, it could be that, for some autosomal genotypes – e.g., (F, G_{ia}) and (M, G_{ia}) – only *women* or only *men* could be in a the subset (G_{is}) – i.e., if either: $P(F, G_{ia}, G_{is}) = 0$ or: $P(M, G_{ia}, G_{is}) = 0$ – whereas, for other autosomal genotypes, both *men* and *women* could be members of the same subset (G_{is}). In such a circumstance, however, MS would then be envisioned as comprising three distinct diseases – one in *men* only, one in *women* only, and a third in both.

Contrasting the Possibilities that: $(\boldsymbol{c = d})$ *or* $(\boldsymbol{c < d})$

When ($\lambda > 0$), to account for an increase in the (*F:M*) *sex ratio*, as shown in *Figure 3*, although it is possible for: ($\boldsymbol{c} = \boldsymbol{d} \leq 1$), this circumstance, nevertheless, seems unlikely. The ascending portion of the response curve is very steep, which indicates that any change in the (*F:M*) *sex ratio* is quite large in response to small changes in exposure. Thus, the window of possible changes in environmental exposure necessary to explain the Canadian data is quite narrow [6]. Moreover, following this narrow window, and for most of these response curves, the (*F:M*) *sex ratio* is declining – a circumstance, which is contrary to evidence [6,22-30]. Also, to achieve an *F:M sex ratio*, which reaches its current level ($i.e.,\ \ p \approx 0.76$), requires either that both the value of (λ) is small and the value of (R) is large (e.g. *Figures 3B–D*), or that the value of (p) is large (e.g. *Figures 3A*). Finally, the increase in failure rate (i.e., the increase in the penetrance of MS for the population), which was observed in Canada between the two *Time Periods*, was large (>32%) and especially prominent among women [3,6]. Each of these circumstances, although still compatible with the condition of: ($\boldsymbol{c} = \boldsymbol{d} = 1$), still seems, nonetheless, to be somewhat at odds with the Canadian data, which demonstrates that the (*F:M*) *sex ratio* has been steadily, and gradually, increasing over many decades [6] and where, currently, the proportion of *women* among MS patients is quite high. By contrast, the circumstances of *Figures 1D, 2D & 4A–D* (i.e., $\boldsymbol{c} < \boldsymbol{d} \leq 1$), result in a continuously increasing (*F:M*) *sex ratio* with increasing exposure over most (or all) of the response curves, they easily account for the magnitudes of the observed (*F:M*) *sex ratios* and, in *Figures 1D & 4A–D*, they could also account for the observation that, at an earlier *Time Point* in the history of MS [40], in both Europe and the United States, the proportion of *men* among individuals with MS seemed to substantially exceed that of *women* such that:

$$0.58 = P(M \mid MS, E_T) > P(F \mid MS, E_T) = 0.42$$

In conclusion, therefore, as indicated in *Methods #4C*, the condition of: ($\boldsymbol{c} < \boldsymbol{d} \leq 1$) is necessarily true for all circumstances, in which ($R \leq 1$) and also, as discussed *above*, seems likely to be true for circumstances in which: ($R > 1$).

<u>*Summary Equations*</u>

For each of the two proportional hazard *Models*, we can use both the observed parameter values, the change in the (*F:M*) *sex-ratio*, and the change in *P*(*MS*) for Canada between any two *Time Periods* [6], and, thereby, construct each of these response curves in their entirety [3]. The values for: Zw_2, Zm_2, Zw_1, Zm_1, I, ($\boldsymbol{d}$), $P(E \mid G,F)$, $P(E \mid G,M)$, and (λ) can then be determined [3] as:

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

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

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

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

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

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

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

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

$$H(a_2) = \ln [1 - Zm_2 / \boldsymbol{c}]$$

$$K(a_2) = R * (H(a_2) - \lambda) = \ln [1 - Zw_2 / \boldsymbol{d}]$$

and: $$\lambda = \{\ln [1 - Zw_2 / \boldsymbol{d}] - \ln [1 - Zm_2 / \boldsymbol{c}]\} / R + [(R - 1)/R] * H(a_2)$$

For the non-proportional *Model*, those parameters, which include ($\boldsymbol{c}$) or ($\boldsymbol{d}$), cannot be estimated from the observed changes in $P(MS \mid F)$ and $P(MS \mid M)$ over time. Notably, the values for $P(F \mid MS)_1$, $P(F \mid MS)_2$, and $P(MS)_2$ have been directly or indirectly observed [3,6]. Also, the values of (Zw_1), (Zw_2), (Zm_1), (Zm_2), $P(E \mid G, F)_2$, $P(E \mid G, M)_2$, ($\boldsymbol{c}$) and ($\boldsymbol{d}$) are, not surprisingly, only related to the circumstances of either *men* and *women*, considered separately. Using a "substitution" analysis, we wrote a computer program, which incorporated the acceptable parameter ranges (*see Methods #2; above*) for the parameters $\{P(G)$; $P(MS \mid MZ_{MS})_2$; $p = P(F \mid G)$; $P(MS)_2$; $P(MS \mid F, G)_2$; $P(F \mid MS)_1$; r; s ; s_a and $C\}$, into the governing equations (*above*) and determined those combinations (i.e., solutions) that fit within the acceptable ranges for both the observed and non-observed parameters (*see Methods #2*). For this analysis, unlike for our *Cross-sectional Model*, we loosened the constraints on the values of (r) and (s) such that:

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

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 $P(G)$ and $P(MS \mid G)_2$ were:

$$0.003 \leq P(G) < 0.55$$
$$0.01 < P(MS \mid G)_2 \leq 0.33$$

If we consider the more restricted range of ($s_a < 1.9$) for the impact of sharing the (E_{twn}) environment with an *MZ*-twin, then:

$$0.003 \leq P(G) < 0.3$$
$$0.05 < P\left(MS \,\middle|\, G\right)_2 \leq 0.3$$

Assuming that the subset (G) does not conform to the <u>*Upper Solution*,</u> but that, considered separately, each of the subsets (F,G) and (M,G) do, and again using *Assertion C*, the range of values for the parameters $P(G), P(MS \mid G)_2, P(MS \mid F,G)_2, P(MS \mid M,G)_2, P(G|F), P(G \mid M)$; and the ratios (x_1/x_2) and (x_1'/x_2') is:

$$0.005 < P(G) \leq 0.82$$
$$0.004 \leq P(G \mid F) \leq 0.66$$
$$0.008 \leq P(G \mid M) \leq 0.99$$
$$0.004 \leq x = P\left(MS \,\middle|\, G\right)_2 \leq 0.12$$
$$0.02 \leq x' = P\left(MS \,\middle|\, IG_{MS}\right)_2 < 0.34$$
$$0.014 \leq x_1 = P\left(MS \,\middle|\, F,G\right)_2 < 0.44$$
$$0.001 \leq x_2 = P\left(MS \,\middle|\, M,G\right)_2 \leq 0.08$$
$$0.03 \leq x_1' = P\left(MS \,\middle|\, F,IG_{MS}\right)_2 < 0.44$$
$$0.002 \leq x_2' = P\left(MS \,\middle|\, M,IG_{MS}\right)_2 < 0.16$$
$$0.004 < P(F \mid G) < 0.66$$
$$1.69 \leq P\left(MS \,\middle|\, F,G\right)/P(MS \mid M,G) = x_1/x_2 \leq 56$$
$$1.86 \leq P(MS \mid F,IG_{MS})/P(MS \mid M,IG_{MS}) = x_1'/x_2' \leq 29$$

We previously concluded that it was possible (or even probable) that men might be disproportionately represented in the subset (G), although any marked disparity in this regard seemed implausible [3]. Therefore, if the restriction of: $\{P(M \mid G) \leq 0.75\}$ is included with the restrictions such that: $(s_a < 1.9)$, $(r \leq 2)$ & $(s \leq 2)$, and also using the "current" *sex-ratio* data from Canada [6] such that: $P\left(F \,\middle|\, MS\right)_2 = 0.74 - 0.78$; (*see Methods #2*), then these estimates are unchanged except that:

$$0.005 < P(G) < 0.3$$
$$0.01 \leq P(G \mid M) < 0.28$$
$$0.07 \leq x_1 = P\left(MS \,\middle|\, F,G\right)_2 \leq 0.31$$
$$0.14 \leq x_1' = P\left(MS \,\middle|\, F,IG_{MS}\right)_2 < 0.43$$
$$0.01 \leq x_2' = P\left(MS \,\middle|\, M,IG_{MS}\right)_2 \leq 0.14$$
$$0.025 \leq P(F|G) \leq 0.66$$
$$1.9 \leq P\left(MS \,\middle|\, F,G\right)/P(MS \mid M,G) = x_1/x_2 \leq 8.3$$
$$2.0 \leq P(MS \mid F,IG_{MS})/P(MS \mid M,IG_{MS}) = x_1'/x_2' \leq 16$$

2. *Longitudinal Model*

Using the *Longitudinal Model*, assuming non-proportional hazards, the possible ranges for these various parameters were:

$$0.001 < P(G) \leq 0.52$$
$$0.001 < P(G \mid F) < 0.32$$
$$0.001 < P(G \mid M) \leq 0.94$$
$$0.10 \leq P(F \mid G) \leq 0.71$$
$$0.004 \leq x = P\left(MS \mid G\right)_2 \leq 0.20$$
$$0.02 \leq x' = P\left(MS \mid G, IG_{MS}\right)_2 \leq 0.34$$
$$0.001 \leq P\left(MS \mid G\right)_1 \leq 0.15$$
$$0.03 \leq x_1 = P\left(MS \mid F, G\right)_2 \leq 0.44$$
$$0.001 < x_2 = P\left(MS \mid M, G\right)_2 \leq 0.125$$
$$0.03 \leq x_1' = P\left(MS \mid F, G, IG_{MS}\right)_2 \leq 0.44$$
$$0.002 \leq x_2' = P\left(MS \mid M, G, IG_{MS}\right)_2 \leq 0.175$$
$$1.2 \leq x_1/x_2 = P\left(MS \mid F, G\right)/P(MS \mid M, G) \leq 32$$
$$2.0 \leq x_1'/x_2' = P(MS \mid F, IG_{MS})/P(MS \mid M, IG_{MS}) < 35$$

In addition, we found that the solution space for both (r) & (s) was restricted: $(r < 20)$ and: $(s < 30)$. Restricting the ranges such that: $(s_a < 1.9)$, $(r \leq 2)$ & $(s \leq 2)$ changes the above estimations such that:

$$0.001 < P(G) < 0.30$$
$$0.001 < P(G \mid F) \leq 0.09$$
$$0.004 < P(G \mid M) \leq 0.54$$
$$0.10 \leq P(F \mid G) \leq 0.65$$
$$0.02 \leq x = P\left(MS \mid G\right)_2 \leq 0.20$$
$$0.10 \leq x' = P\left(MS \mid G, IG_{MS}\right)_2 \leq 0.34$$
$$0.005 \leq P\left(MS \mid G\right)_1 \leq 0.15$$
$$0.10 \leq x_1 = P\left(MS \mid F, G\right)_2 \leq 0.30$$
$$0.006 < x_2 = P\left(MS \mid M, G\right)_2 \leq 0.125$$
$$0.14 \leq x_1' = P\left(MS \mid F, G, IG_{MS}\right)_2 \leq 0.42$$
$$0.008 \leq x_2' = P\left(MS \mid M, G, IG_{MS}\right)_2 \leq 0.175$$
$$1.7 \leq x_1/x_2 = P\left(MS \mid F, G\right)/P(MS \mid M, G) \leq 27$$
$$2.0 \leq x_1'/x_2' = P(MS \mid F, IG_{MS})/P(MS \mid M, IG_{MS}) < 27$$

Using the *Longitudinal Model*, assuming ($\boldsymbol{c} = \boldsymbol{d} = 1$) and, thus, with ($R = R^{app} = q_w^{min}/q^{min}$), the possible ranges for these parameters are unchanged from the unrestricted non-proportional *Model* except for the additional estimations of:

$$0.0005 \leq \lambda \leq 0.13$$
$$1.3 \leq R^{app} \leq 1177$$
$$0.03 \leq P(E \mid F, G) \leq 0.44$$
$$0.001 \leq P(E \mid M, G) \leq 0.125$$
$$1.2 \leq P(E \mid F, G)/P(E \mid M, G) \leq 32$$

If the solution space were restricted: such that: ($s_a < 1.9$), ($r \leq 2$) & ($s \leq 2$), the above estimates are unchanged except that:

$$1.9 \leq R^{app} \leq 516$$
$$0.10 \leq P(E \mid F, G) \leq 0.30$$
$$0.006 \leq P(E \mid M, G) \leq 0.125$$
$$1.9 \leq P(E \mid F, G)/P(E \mid M, G) \leq 27$$

If ($R = 1$), these estimates are the unchanged from the non-restricted values above except:

$$0.002 < \lambda < 2.4$$
$$\boldsymbol{d} = 1 \quad \textit{(by Assumption – see Methods \#4F)}$$
$$0.002 \leq \boldsymbol{c} \leq 0.79$$
$$1.3 < \boldsymbol{d}/\boldsymbol{c} < 493$$
$$0.03 \leq P(E \mid F, G) \leq 0.3$$
$$0.03 \leq P(E \mid M, G) < 0.94$$
$$0.04 \leq P(E \mid F, G)/P(E \mid M, G) \leq 0.95$$

As in our analysis of the *Cross-sectional Model* (above), if the restriction of: $\{P(M \mid G) \leq 0.75\}$ is included with the above restrictions such that: ($s_a < 1.9$), ($r \leq 2$) & ($s \leq 2$) then, for ($R = 1$), these estimates become:

$$0.001 < P(G) < 0.10$$
$$0.001 < P(G \mid F) \leq 0.09$$
$$0.001 < PG \mid M) \leq 0.08$$
$$0.08 \leq x = P(MS \mid G)_2 \leq 0.20$$
$$0.10 \leq x' = P(MS \mid G, IG_{MS})_2 \leq 0.34$$
$$0.02 \leq P(MS \mid G)_1 \leq 0.15$$
$$0.10 \leq x_1 = P(MS \mid F, G)_2 \leq 0.30$$
$$0.03 < x_2 = P(MS \mid M, G)_2 \leq 0.125$$
$$0.14 \leq x_1' = P(MS \mid F, G, IG_{MS})_2 \leq 0.42$$

$$0.03 \leq x_2' = P\left(MS \,\middle|\, M, G, IG_{MS}\right)_2 \leq 0.175$$
$$1.7 \leq x_1/x_2 = P\left(MS \,\middle|\, F, G\right)/P(MS \,|\, M, G) \leq 4.5$$
$$2.0 \leq x_1'/x_2' = P(MS \,|\, F, IG_{MS})/P(MS \,|\, M, IG_{MS}) < 8.6$$

$$0.007 < \lambda < 2.4$$
$$0.04 \leq \boldsymbol{c} \leq 0.55$$
$$1.8 < \boldsymbol{d}/\boldsymbol{c} < 26$$
$$0.1 \leq P\left(E \,\middle|\, F, G\right) \leq 0.3$$
$$0.11 \leq P\left(E \,\middle|\, M, G\right) \leq 0.93$$
$$0.18 \leq P\left(E \,\middle|\, F, G\right)/P\left(E \,\middle|\, M, G\right) \leq 0.95$$

Discussion

The present analysis provides considerable insight to the nature of susceptibility to MS. Thus, both of our *Models,* and the intersection of all our analyses, substantially support each other. For example, regardless of the whether the *Cross-sectional* or the *Longitudinal Model* was used, regardless of the whether the hazards are proportional and, if proportional, regardless of the proportional *Model* assumed, the consistently supported range for $P(G)$ is:

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

Thus, under any circumstance, a large percentage of the general population ($\geq 48\%$), and likely the majority, must be impervious to getting MS, regardless of their environmental experiences. Consequently, if a person doesn't have the appropriate genotype, they can't get the disease. This conclusion is particularly evident for *women*, where:

$$0.001 < P\left(G \,\middle|\, F\right) \leq 0.30$$

Thus, much of the population and most *women* lack this essential component of MS pathogenesis. In this sense, MS is fundamentally a genetic disorder.

Nevertheless, MS is also, fundamentally, an environmental disease. Thus, over the last several decades, the prevalence (and, thus, the penetrance) of MS has increased in many parts of the world [6,22-30]. Because genetic factors do not change this quickly, this fact implicates environmental factors as also critical to disease pathogenesis [3]. This conclusion is also supported by the observation that (E_{twn}) environment seems to significantly impact the likelihood that an individual either has, or will subsequently develop, MS [7,31-37].

In considering the observations that have been made for *men* and *women* in Canada [6], one possible explanation is that the hazard functions for developing MS in the two genders are not

proportional. In this view, each gender develops disease in response to distinct sets of environmental events (*see Methods #4A*) and, thus, MS in *women* represents a fundamentally different disease than MS in *men*. Moreover, in this non-proportional view, the environmental changes, which have taken place in Canada between the two *Time Periods* of 1941-1945 & 1975-1980 (whatever these are), would be interpreted as involving those events that impact MS development in susceptible *women* to a considerably greater extent than they do those events that impact MS development in susceptible *men*. However, even in this case, the limits derived for the parameters $\{(PG), P(G \mid F), x, x_1, x_2, x_1', x_2', x_1/x_2$ and $x_1'/x_2'\}$ would still apply.

Nevertheless, the notion that MS in *men* and MS in *women* are fundamentally different diseases, involving distinct environmental events, seems implausible. Importantly, the view that hazards are proportional does not depend upon every individual (or *i-type*) having either the same proportionality constant or the same the threshold difference (*see Methods #4F*). Rather, it depends only upon the same environmental events having a non-zero probability of impacting the development of MS in both susceptible *women* and *men* (*see Methods #4A*). It is of note, therefore, that both genders seem to share very similar mechanisms of disease pathogenesis. Indeed, there have been several epidemiological observations that link MS, unequivocally, to environmental factors, to genetic factors, or to both and when these have been explored systematically, these factors seem to impact both *men* and *women* in a similar manner. For example, a *month-of-birth* effect has been reported in MS whereby, in the northern hemisphere, the risk of subsequently developing MS is greatest for babies born in May and least for babies born in November compared to other months during the year [41]. This *month-of-birth* effect was predicted to be inverted in the southern hemisphere [41] and, in fact, a subsequent *population-based* study from Australia found the peak risk to be for babies born in November-December and the nadir to be for babies born in May-June [42]. Although this *month-of-birth* effect is somewhat controversial [43], it has been widely (and reproducibly) reported by many authors and the effect is apparent in both *men* and *women* [41,42,44,45]. Thus, MS-risk seems to cycle throughout the year and this observation, if correct, clearly, implicates an environmental factor (or factors) – affecting both *men* and *women* alike – which is (are) linked to the solar cycle and occur(s) during the intrauterine or early post-natal period [8]. Second, the recurrence risk of MS is generally found to be greater in a co-twin of a *DZ*-twin proband with MS compared to a non-twin co-sibling of a sibling proband with MS [7,31-37]. This effect also implicates an environmental factor that occurs in proximity to the birth and is apparent in both *men* and *women* [3,8]. Third, it is widely reported that MS becomes increasingly prevalent in those geographic regions, which lie farther (either north or south) from the equator [8,46,47]. This observation could implicate either environmental or genetic factors but, regardless, this gradient is apparent in both *women* and *men* [46,47]. Fourth, evidence of a prior *EBV* infection is found in essentially all MS patients [8,48]. If this infection is

actually present in 100% of MS patients, then this indicates that an *EBV* infection is a necessary factor in the causal pathway leading to MS for all susceptible *women* and *men* [8]. Fifth, vitamin D deficiency has been implicated as being an environmental factor in MS pathogenesis [8,49-53] and this factor is related to MS in both *men* and *women* [49-52]. And lastly, smoking tobacco has been implicated as being environmental factor associated with MS pathogenesis [8,54] and, again, this factor is associated with MS in both *women* and *men* [54].

Also, the genetic basis of MS seems to be very similar in both *women* and *men*. Thus, the strongest genetic associations with MS are for certain haplotypes within the *HLA*-region on the short arm of Chromosome 6 [55-59] and, in the predominantly Caucasian population of the Wellcome Trust Case Control Consortium (*WTCCC*) dataset [60,61], the most strongly MS-associated haplotypes in this region are similarly associated with MS in both *women* and *men* (*Tables 3 & 4*).

In addition, it is useful to further consider the notion that, to explain the threshold difference, susceptible *women* might require a more *"intense"* environmental exposure than susceptible *men*. For example, in *Methods #4F*, we entertained the notion that different *i-types* might potentially require different families of sufficient environmental exposure $\{E_i\}$, each with a potentially different threshold difference (λ_i), and each with a potentially different proportionality constant (R_i) and, potentially, with many different sufficient sets within each family. It seems rather surprising, if such marked variability truly existed, that this could possibly lead to a circumstance in which <u>*all*</u> susceptible *women* required a more "*intense*" exposure compared to <u>*some*</u> susceptible *men*. By contrast, if all *i-type* individuals required the same set (or sets) of environmental exposure, a consistent difference in the *"intensity"* of the required exposures might be easier to rationalize (see our earlier discussion regarding the possible role that *EBV* and Vitamin D deficiency might have in MS pathogenesis – *Methods #4F*). Also, if the different *i-types* require the same set (or sets) of environmental exposure, this also makes it easier to rationalize the fact that those environmental factors, which have been consistently identified as MS-associated, have been linked to MS, generally, but not to any particular subgroup [8,41,42,44-54].

As noted earlier (*see Methods #4A & 4F; Equations 7 & 12b*), by assuming that: ($\boldsymbol{c} = \boldsymbol{d} \leq 1$), we are also assuming that the difference in disease expression between *men* and *women* is due entirely to a difference in the likelihood of their experiencing a sufficient environmental exposure. Therefore, because currently ($Zw_2 > Zm_2$) – *see Methods #2C* – those conditions in which ($\boldsymbol{c} = \boldsymbol{d} \leq 1$) would necessarily lead to the conclusion that, currently, susceptible *women* are more likely to experience a sufficient environment compared to susceptible *men* despite the fact that the probabilities of exposure to each family, $P(\{E_i\} \mid E_T)$ and $P(\{E_{iw}\} \mid E_T)$, are fixed constants during any (E_T).

Moreover, there are also several additional lines of evidence, which, taken together, strongly

suggest that the circumstance of $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$ is unlikely. First, for all $(R \leq 1)$, it must be the case that: $(\boldsymbol{c} < \boldsymbol{d} \leq 1)$ – *see Methods #4D & 4E*. Second, as discussed earlier (*see Methods #4F*), the response curves required for conditions where: $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$ & $(R > 1)$ have very steep ascending portions and, thus, present only a narrow window of opportunity to explain the Canadian data [6] regarding the changes in the (*F:M*) *sex ratio* and its magnitude over time (*see Methods #4F; see also Figure 3*). Also, for these response curves, following this narrow window, and contrary to evidence [6], the (*F:M*) *sex ratio* decreases with increasing exposure (*Figure 3*). By contrast, these Canadian data suggest that there has been a gradual and sustained increase in the (*F:M*) *sex ratio* over the past several decades [6]. Third, as noted in *Methods #1D*, there seems to be little impact of the (E_{sib}) environment on the development of MS. However, when $(\boldsymbol{c} = \boldsymbol{d})$ the only explanation for $(R > 1)$ is a disproportionate exposure to sufficient environments experienced by *women* (*see above*). Thus, proband siblings and their non-twin co-siblings (both *men* and *women*), despite sharing common genes and a common childhood environment, still depend upon (and differ in) only their (E_{pop}) exposures to develop their MS. Fourth, if $(\lambda > 0)$, as it must be for all for $(R \geq 1)$, there is an inherent tension between the fact that *men* are responding to a broader range of environmental exposures (i.e., to both the less and the more *intense* exposures) compared to *women* (who respond to only the more *intense* exposures) and, yet, that *women* are more likely to experience a sufficient environmental exposure compared to *men*.

Fifth, and most importantly, for each of the known (or suspected) environmental factors related to MS pathogenesis, there is no evidence to suggest that *women* are disproportionately experiencing them. Thus, the *month-of-birth* effect is equally evident for *men* and *women* [41,42,44,45]; the latitude gradient is the same for both genders [8,46,47]; the impact of the (E_{twn}) environment is the same for *men* and *women* [3]; By young adulthood (i.e., 20-25 years), the likelihood of an *EBV* infection (a factor probably in the causal chain leading to MS), is about equal (~95%) for both genders, although infection likely occurs earlier among *women* [8,62,63]; vitamin D levels are the same in both genders [49-53]; and smoking tobacco is actually more common among *men* [8,54]. Taken together, these epidemiological observations suggest that *women* and *men* are currently experiencing the same relevant environmental events in an equivalent manner. Therefore, each of these observations suggests that:

$$P(E \mid G, F, E_T)_2 \approx P(E \mid G, M, E_T)_2 \approx P(E \mid G, E_T)_2 \qquad \textit{(Equation 13a)}$$

In this context, the possibility that $(\boldsymbol{c} = \boldsymbol{d} \leq 1)$ & $(R > 1)$ – which are noted in the *Results* and are depicted in *Figure 3* – seems remote, especially given the facts that the relevant exposures are *population-wide* and that the difference between (Zw) and (Zm) can only be explained by a disproportionate exposure to sufficient environments by susceptible *women* (*see Equation 10a*) – a circumstance for which there is decidedly no evidence (*see above; see also, Equation 12b*). Therefore, if

Equation 13a is correct, then also:

$$P(E_w \mid G, F, E_T)_2 \approx P(E_w \mid G, M, E_T)_2 \approx P(E_w \mid G, E_T)_2 \qquad (\textit{Equation 13b})$$

Moreover, because, the population experiences the same level of exposure $(u = a)$ during any (E_T), therefore, from *Equations 6* and *9*, these approximate equivalences (if correct) would indicate both that: $(\boldsymbol{c} < \boldsymbol{d} \leq 1)$ and that: $(R \approx 1)$ – *see Methods #4F*. Such a configuration easily explains an increasing (*F:M*) *sex ratio* and its magnitude throughout most (or all) of the response curves (*see Figures 4A–D*), it accounts for a time in MS history where the disease may have been more prevalent in *men* [40] – e.g., *Figures 4C & 4D* – and it does not assume that, even though susceptible *men* and *women* have the same *population-wide* exposures, $\{P(\{E_i\} \mid E_T)\}$ and $\{P(\{E_{iw}\} \mid E_T)\}$, there is, nonetheless, a marked and systematic increase in the exposure to (E_w) for susceptible *women* compared to *men.* However, any condition, for which $(\boldsymbol{c} < \boldsymbol{d})$, does require that some susceptible *men* will never develop MS, even when the correct genetic background occurs together with an environmental exposure sufficient to cause MS in those individuals. Indeed, if, as suggested: $(R = 1)$, then, necessarily, $(\boldsymbol{c} < \boldsymbol{d})$ – *see Methods #4F, above* – and, indeed, such *men* will comprise 22-98% of the susceptible *male* subset (M, G). Naturally, in such a circumstance, it seems likely that the proportion of *women* who ultimately develop MS, given the same conditions, will also be less than unity (e.g., *Figures 4B & 4D*). However, because we needed to assume that: $(\boldsymbol{d} = 1)$ for the purposes of our analysis, this possibility cannot be addressed using the Canadian data.

Some of the individuals who don't seem to develop MS despite having an environmental exposure sufficient to cause MS, no doubt, will have subclinical disease. Indeed, as suggested by several autopsy studies, the prevalence of "asymptomatic" MS in the population (Z) may be as high as ~0.1% [64-67]. Moreover, such a figure is generally supported by several magnetic resonance imaging (*MRI*) studies of asymptomatic individuals [68,69]. Nevertheless, although these considerations suggest that some proportion of MS can be asymptomatic, this fact seems unlikely to account for any difference either (***c***) from (***d***), or of (***c***) from the expected 100% occurrence of MS in *men* who are both genetically susceptible and, in addition, experience an environment sufficient to cause MS given their specific genotype. Thus, if asymptomatic disease did account for (***c***) being less than (***d***), then *men* should account for a disproportionately large percentage of these asymptomatic individuals. However, this is decidedly not the case. Rather, *men* account for only 16% of the asymptomatic individuals detected by *MRI* [68,69] – a percentage well below their proportion of symptomatic cases [3]. Consequently, if $(\boldsymbol{c} < \boldsymbol{d} \leq 1)$, as the Canadian data [6] seems to indicate, then chance must play a role in disease pathogenesis.

Alternatively, however, perhaps our previous definition of exposure "*intensity*" does not account properly for certain other potential aspects of exposure "*intensity*", which might play an important role in

disease pathogenesis. As a concrete example of this notion, suppose, as before, that one (or more) of the sufficient sets of exposures for the ith susceptible individual includes both a deficiency of vitamin D and a prior *EBV* infection [3], each occurring during some critical period of the person's life (not necessarily during the same period). Furthermore, suppose that, with all other necessary factors being equal in the ith susceptible individual, a mild vitamin D deficiency for a short period during the critical time, together with an asymptomatic *EBV* infection at age 10, causes MS to develop 10% of the time, whereas a more prolonged, and more marked, vitamin D deficiency during the critical period, together with a symptomatic *EBV* infection (mononucleosis) at age 15, causes MS to develop 75% of the time. Notably, each of these posited conditions is sufficient, by itself, to cause MS; the only difference is in the likelihood of this outcome, given the different levels (i.e., "*intensity*") of exposure. Although this notion of "*intensity*" differs from our previous definition and can't be easily quantified, presumably, there will be a positive correlation between an increasing "*intensity*" of this exposure (whatever this means operationally) and an increasing risk of MS for each susceptible individual. Moreover, each susceptible individual must reach a *maximum* likelihood of developing MS as the "*intensity*" of their exposure increases. This maximum may be at 100% or it may be at something less than this but, whatever it is, there must be a maximum for each person. In addition, unlike our previous definition of $P\{E_i\}$, where only one sufficient set of exposures was necessary, here, an individual for whom two or more of their sufficient sets of exposure occur, may experience a greater "*intensity*" of exposure than if only one set occurs. Nevertheless, none of these circumstances alters the fact that each person will still have their "*maximum*" likelihood of developing MS under optimal environmental conditions. We can then define the "*intensity*" of exposure – $P(E \mid G, E_T)$ – as the average (or expected) "*intensity*" of exposure (however this is measured) experienced by members of the (G) subset, given the environmental conditions of the time (E_T). When no sufficient exposure occurs for any member of (G): $P(E \mid G, E_T) = 0$. When the "*intensity*" has increased to the point where every member of (G) has reached their maximum likelihood of developing MS then: $P(E \mid G, E_T) = 1$. And, again, we can define (u), as the odds of exposure:

$$u = P(E \mid G, E_T)/[1 - P(E \mid G, E_T)].$$

Although, clearly, this conceptualization of exposure *intensity* is different (and perhaps more realistic) than the "*sufficient*" exposures considered earlier, two of its features are particularly noteworthy. First, randomness is integral to this notion of exposure "*intensity*". Thus, disease expression at low "*intensity*" exposures, by definition, incorporates an element of chance because the likelihood of developing MS under these conditions must be less than the *maximum*, for at least some susceptible individuals. If not, then this "*intensity*" of exposure would have no impact on anyone, and this *Model* becomes equivalent to the "*sufficient*" exposures *Model* considered earlier. Second, despite exposure

being measured differently, and despite the hazard functions likely being different, all the equations and transformations presented in *Methods #4A–F* (*above*) as well as the calculated response curves are unchanged by measuring exposure as *"intensity"* in this manner rather than as *"sufficiency"*. Indeed, this conclusion applies to any measure of exposure, which incorporates the notion of "sufficient sets" of environmental exposure defined earlier (*see Methods #1B*).

Thus, by any measure, the Canadian data [6] seem to indicate that there is a stochastic factor (i.e., an element of chance) in MS pathogenesis, at least for *men*, which determines, in part, who gets the disease and who does not. Such a conclusion might be viewed as surprising because, in the universe envisioned by many physicists, events are (or seem to be) deterministic. For example, imagine a rock thrown at a window. If the rock has a mass, a velocity, and an angle of impact sufficient to break the window, given the physical state of the window at the moment of impact, then we expect the window to break 100% of the time. If the window only breaks some of the time, likely, we would conclude that we hadn't adequately specified the sufficient (i.e., initial) conditions. If the *population-based* observations in over 29,000 Canadian MS patients are to be believed, however, this is not so for the development of MS. Even for an individual with a susceptible genotype and an environmental experience sufficient to cause disease given their specific genotype, they still may or may not develop the illness. This result cannot be ascribed to contributions from other, unidentified, environmental factors because each set of environmental circumstances considered here is defined to be sufficient, by itself, to cause MS in that specific susceptible individual. If other environmental conditions were needed to cause MS reliably in that individual, these conditions would already be necessary components of these sufficient environments. Even altering the definition of exposure to include the importance of different *intensities* of exposure doesn't alter this conclusion. Certainly, the invocation of stochastic processes in disease pathogenesis requires replication, both in MS and in other disease states, before being accepted as fact. Nevertheless, if replicated, such a result would imply that there is a fundamental randomness to the behavior of some complex physical systems (e.g., organisms), in contrast to the apparent determinism of physical laws [70].

It is of note that some authorities have argued (albeit controversially) from fundamental physical principles that true randomness (i.e., thermodynamic equilibrium or maximum entropy) was a primordial property of our universe in the earliest tiny fraction of a second of the big bang and that this inherent randomness is reflected by a currently observable randomness for both microscopic (i.e., quantum uncertainty) and macroscopic descriptions of the universe [71]. By contrast, the deterministic hypothesis envisions that earliest state of the universe was one of minimum entropy and asserts that, when we perceive certain macroscopic events as being due to chance, this perception is illusory and merely a reflection of our ignorance regarding the relevant initial conditions [70,71]. One author has stated this deterministic point of view succinctly by noting that, while "the quantum equations lay out many possible

futures, … they deterministically chisel the likelihood of each in mathematical stone" [70]. Obviously, the question of which, if either, of these alternate views of the universe represents reality has far-reaching implications [70,71]. Perhaps the best contemporary evidence for macroscopic randomness, cited by proponents of the non-deterministic worldview, is the case of biological evolution by means of natural selection [71]. Thus, natural selection is envisioned to be a non-sentient process, which depends upon the occurrence of apparently random events and, using these events, permits living species to respond both continuously and adaptively to the varying environmental conditions of different times, or different places, or both. Moreover, the direction, in which any new species evolves, is seemingly not predictable but, rather, depends upon the nature of the specific random events, which take place.

Placed into a broader context, this biological evolution, which has been so clearly documented on Earth, is probably best viewed as a part of (or as a continuation of) the process of chemical evolution – a process that began only a few minutes after the onset of the big bang, and at a time when the universe was composed of 75% hydrogen, 25% helium, and a small admixture of lithium [70,72]. The chemistry of this early universe was extremely rudimentary. Helium is a nobel gas and reacts with essentially nothing. Hydrogen and lithium combine to form only a few simple chemical compounds such as lithium hydride (LiH) and molecular hydrogen (H_2). A more complex chemistry (and, in particular, the chemistry necessary to create and sustain life and, also, to permit biological evolution) only evolved later with synthesis of the heavier atomic elements – a synthesis that, following these first few minutes of the big bang, only occurred with the collapse and/or explosion of massive stars at the end of their life cycle [70,72]. This synthesis, and the subsequent build-up of heavier elements in the universe, was gradual and took time. Also, this process of chemical evolution continues to this day, not only with the ongoing synthesis of heavier elements inside contemporary stars and the interactions of these elements with each other throughout the universe, but also with the synthesis of a multitude of novel chemical compounds, created by living organisms (including humans). Moreover, each step of this evolutionary sequence seems to require the occurrence of random events – i.e., which nuclei happen to collide, whether they fuse, whether (and when) they decay, which star becomes a supernova, where and when these supernovas occur, which life-forms evolve, and under what circumstances, with what chemistries, and in what places, etc.

In this broader context, then, it is hard to imagine that the past and future course of this evolutionary sequence was, and is, a pre-determined outcome and yet, in the case of biological evolution on Earth, for the macroscopic process that produces it, both to be so exquisitely adaptive to contemporary external events and yet, to be so completely dependent upon apparently random occurrences. Nevertheless, it is extremely difficult to prove that any macroscopic process (including this evolutionary sequence) is truly random. Despite this difficulty, however, the hypothesis of determinism is quite fragile

in the sense that, if the true randomness of even one macroscopic process or event could be established, the hypotheses of determinism would be undermined. Consequently, if replicated, any demonstrated randomness in MS disease expression or in the expression of any other disease {i.e., any circumstance, in which either: $(\boldsymbol{c} < \boldsymbol{d} = 1)$ or: $(\boldsymbol{c} \leq \boldsymbol{d} < 1)$}, would provide strong empiric evidence in support for the non-deterministic worldview. By contrast, the deterministic worldview requires the condition that: $(\boldsymbol{c} = \boldsymbol{d} = 1)$.

There are two features of the response curves in *men* and *women* that merit further comment. First, the plateaus for these curves (if, in fact, $\boldsymbol{c} < \boldsymbol{d} \leq 1$) – e.g., *Figures 1D, 2D, 4A–D* – reflect this inherent randomness in the process of disease development. Indeed, in this circumstance, it would be this randomness, rather than the genetic and environmental determinants, which lies at the heart of the difference in disease expression between *men* and *women*. Thus, genetically susceptible *women*, who experience an environment sufficient to cause MS given their genotype, are more likely to develop disease compared to susceptible *men* in similar circumstances. Consequently, if $(\boldsymbol{c} < \boldsymbol{d})$, there must be something about "*female-ness*" that favors disease development in *women* over *men* although, whatever this is, it is not part of any causal chain of events leading to disease (i.e., in the sense that, if a coin-flip determines, in part, an outcome, then this random event is not part of any <u>*causal*</u> chain). As noted above, if either $(\boldsymbol{c} < \boldsymbol{d} = 1)$ or: $(\boldsymbol{c} \leq \boldsymbol{d} < 1)$, then disease development in the setting of a susceptible individual experiencing a sufficient exposure must include a truly random event (at least for *men*). Moreover, if: $(\boldsymbol{c} < \boldsymbol{d})$, the fact that this random process favors disease development in *women* does not make it any less random. For example, the flip of a biased coin is no less random than the flip of a fair coin. The only difference is that, in the former circumstance, the two possible outcomes are not equally likely. In the context of MS, "*female-ness*" would then be envisioned to bias the coin differently than does "*male-ness*" (whatever these terms mean).

Second, the thresholds reflect the minimum exposure at which disease expression begins and the response curves, with increasing exposure, that follow this onset [3], need to account for both the increasing prevalence of MS and the steadily increasing proportion of *women* among MS patients [6,22-30,40]. If the hazards in *men* and *women* are not proportional, as discussed *above*, little accounting is necessary. By contrast, if the hazards are proportional and if both: $(\boldsymbol{c} < \boldsymbol{d} \leq 1)$ & $(\lambda > 0)$, then this could account for all of the epidemiological observations – i.e., the increasing prevalence of MS [22-30], the continuously increasing proportion of *women* among MS patients [6,22-30,40], the magnitudes of the observed (*F:M*) *sex ratios* [3,6], and a 1922 study [3,40], in which MS prevalence in both the United States and Europe was reported to be substantially higher in *men* than in *women* (e.g., *Figures 1D, 4A–D*).

During the development of our *Longitudinal Model*, we observed that when the prevailing

environmental conditions of a time (E_T) were such that: $\{P(E \mid E_T) = 0\}$, 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 [73,74]. Perhaps, the earliest described case of MS was that of Saint Lidwina of Schiedam (c. 1396) although the argument that Augustus d'Este (c. 1822) suffered from MS seems more unequivocal [73]. Even though 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, if MS existed, case descriptions (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, ever since its original description, MS seems to be changing in character – a fact that underscores the critical importance of environmental factors in MS pathogenesis. For example, although considered uncommon initially, ever since Charcot's initial characterization, MS has become increasingly recognized as a common neurological condition [74-76]. 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 [74-76]. Moreover, in the late 19th and early 20th centuries, the disease was thought to be more (or equally) prevalent in *men* compared to *women* [40,74,75], whereas, today, *women* account for 66–76% of the cases [3,76]. Also, in many parts of the world, MS is increasing in frequency, particularly among *women* [6,22-30]. Indeed, in Canada, $P(MS)$ has increased by an estimated minimum of 32% over a span of 35–40 years [3] – a circumstance which has led to a 10% increase in the proportion of *women* among MS patients ($p < 10^{-6}$) over the same time-interval [6].

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 (e.g., *Table 3*), 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* [55-59]. Each of these haplotypes, as well as each of the conserved extended haplotypes (*CEHs*) – *see Table 4* – is well represented in diverse human populations around the globe [77,78] 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 epidemiology, which have taken place over the past two centuries. Consequently, it seems that $\{P(E \mid G, E_T) = 0\}$ is possible under those environmental conditions that existed prior to the late 14th century and, thus, that "purely genetic" MS does not exist.

In summary, the development of MS (in an individual) requires both that they have an appropriate genotype (which is uncommon in the population) and that they have an environmental exposure sufficient to cause MS given their individual genotype. Nevertheless, even when the necessary genetic and environmental factors, required for MS pathogenesis, co-occur for an individual, this still seems to be insufficient for that person to develop MS. Thus, disease pathogenesis, even in this circumstance, seems not to be deterministic but, rather, to involve an important element of chance.

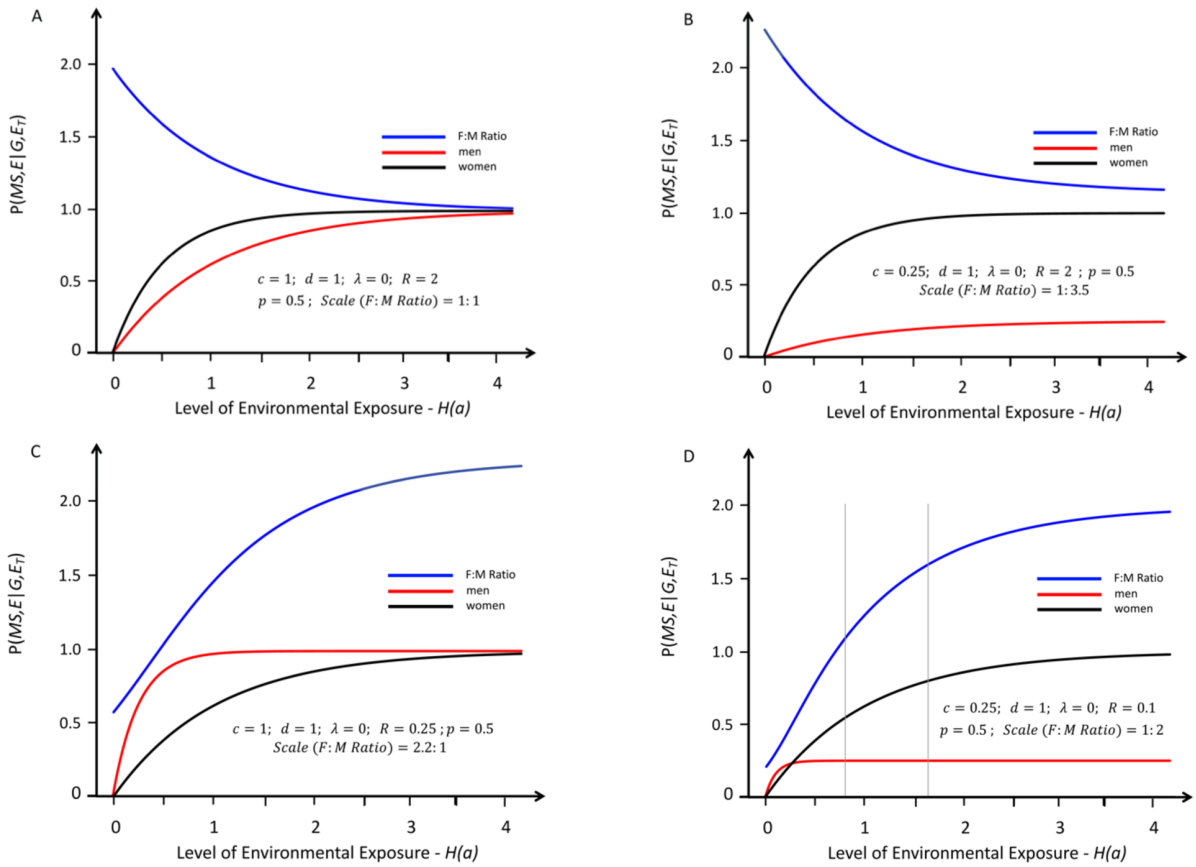

Figure 1. Response curves representing the likelihood of developing MS in genetically susceptible *women* (black lines) and *men* (red lines) with an increasing probability of a "sufficient" environmental exposure – *see Methods #1B*. The curves depicted are "strictly" proportional, meaning that the environmental threshold is the same for both *men* and *women* – i.e., under conditions in which: ($\lambda = 0$) – *see Text*. The blue lines represent the change in the (*F:M*) *sex ratio* (plotted at various scales) with increasing exposure. The thin grey vertical lines represent the portion of the response curve that covers the change in the (*F:M*) *sex ratio* from 2.2 to 3.2 (i.e., the actual change observed in Canada [6] between *Time Periods #1 & #2*). The grey lines are omitted under circumstances either where these observed (*F:M*) *sex ratios* are not possible or where both ($Zw > Zm$) and an increasing (*F:M*) *sex ratio* are not possible. Response curves *A* and *B* reflect conditions in which ($R > 1$); whereas curves *C* and *D* reflect conditions in which ($R < 1$). If ($R = 1$), the blue line would be flat. Response curves *A* and *C* reflect conditions in which ($\boldsymbol{c} = \boldsymbol{d} = 1$); whereas curves *B* and *D* reflect those conditions in which ($\boldsymbol{c} < \boldsymbol{d} = 1$). Under the conditions for curves *A* and *B* ($R \geq 1$), there is no possibility that the (*F:M*) *sex ratio* will be observed to increase with increasing exposure. Under the conditions of curve *C* – i.e., ($\boldsymbol{c} = \boldsymbol{d} = 1$) and ($R < 1$) – at no exposure level is it possible that:

$$Zw = P(MS, E \mid G, F, E_T) > P(MS, E \mid G, M, E_T) = Zm$$

Thus, the only "strictly" proportional model that could possibly account for an increasing (*F:M*) *sex ratio*, and for the fact that: ($Zw_2 > Zm_2$), is a *Model* in which ($\boldsymbol{c} < \boldsymbol{d} \leq 1$) and ($R < 1$) – i.e., curve *D*.

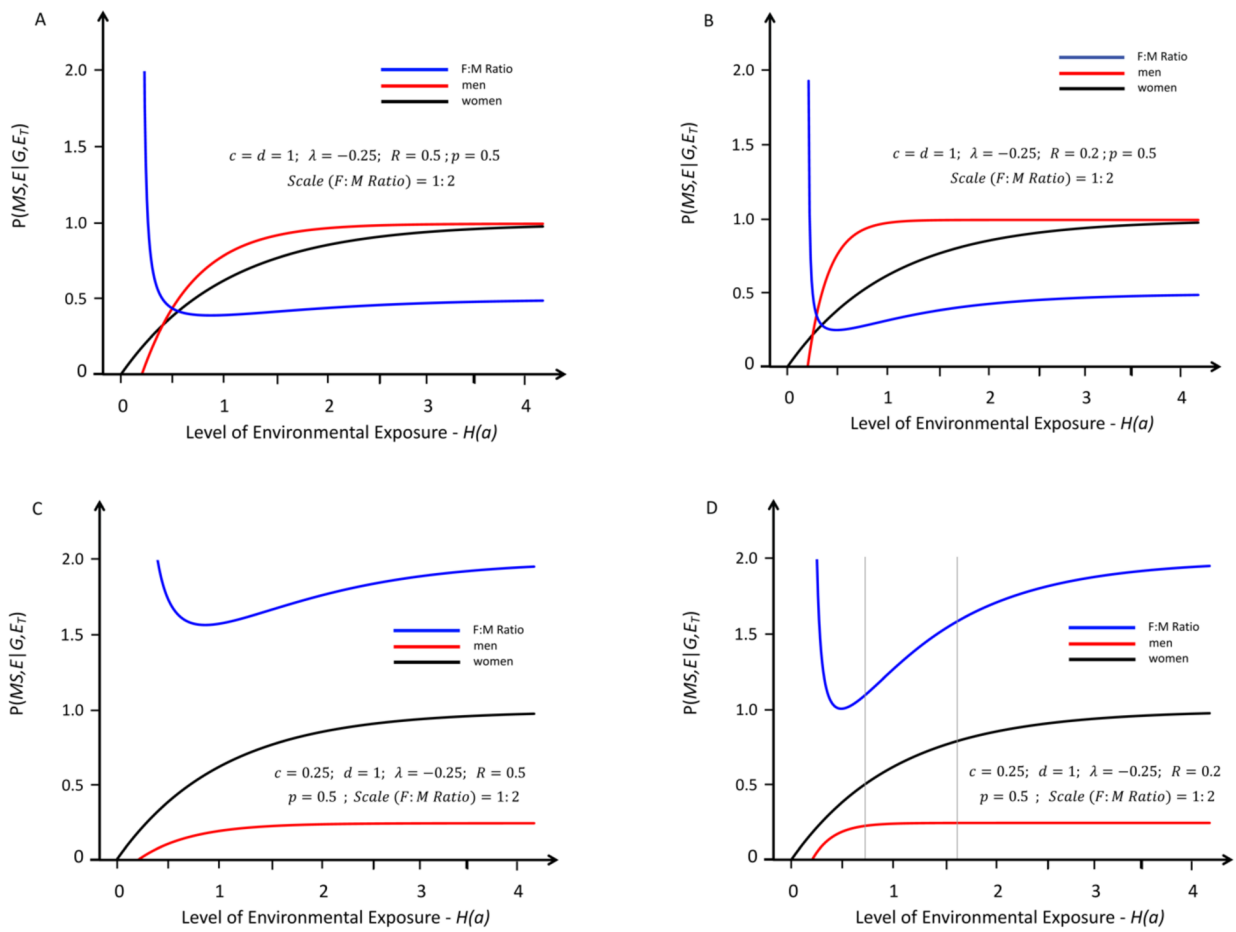

Figure 2. Response curves for the likelihood of developing MS in genetically susceptible *women* (black lines) and *men* (red lines) with an increasing probability of a "sufficient" environmental exposure – *see Methods #1B*. Like *Figure 1*, the curves depicted are also proportional although here the environmental threshold is greater for *men* than for *women* – i.e., under conditions in which: $(\lambda < 0)$ – *see Text*. The blue lines represent the change in the (*F:M*) *sex ratio* (plotted at various scales) with increasing exposure. The thin grey vertical lines represent the portion of the response curve that covers the change in the (*F:M*) *sex ratio* from 2.2 to 3.2 (i.e., the actual change observed in Canada [6] between *Time Periods #1 & #2*). The grey lines are omitted under circumstances either where these observed (*F:M*) *sex ratios* are not possible or where both an increasing (*F:M*) *sex ratio* and $(Zw > Zm)$ are not possible. Response curves *A* reflects conditions in which $(\boldsymbol{c} = \boldsymbol{d} = 1)$ & $(R > 1)$; Response curves *B* reflects conditions in which $(\boldsymbol{c} = \boldsymbol{d} = 1)$ & $(R < 1)$; curve *C* reflects conditions in which $(\boldsymbol{c} < \boldsymbol{d} = 1)$ and $(R < 1)$ and curve *D* reflects those conditions in which $(\boldsymbol{c} < \boldsymbol{d} = 1)$ and $(R < 0.5)$. To account for the observed increase in the (*F:M*) *sex ratio*, curve *D* (compared to curve *C*) requires a small enough value of (R) so that the (*F:M*) *sex ratio* curve dips below 2.2 and, also, a small enough value of $(\boldsymbol{c})$ so that the curve rises above 3.2. For all points for curve *A* $(Zw/Zm > \boldsymbol{d}/\boldsymbol{c})$. Curves *B* & *C* never even approach the (*F:M*) *sex ratio* of 2.2. By contrast, for curve *D*, an appropriate increase in the (*F:M*) *sex ratio* can be observed.

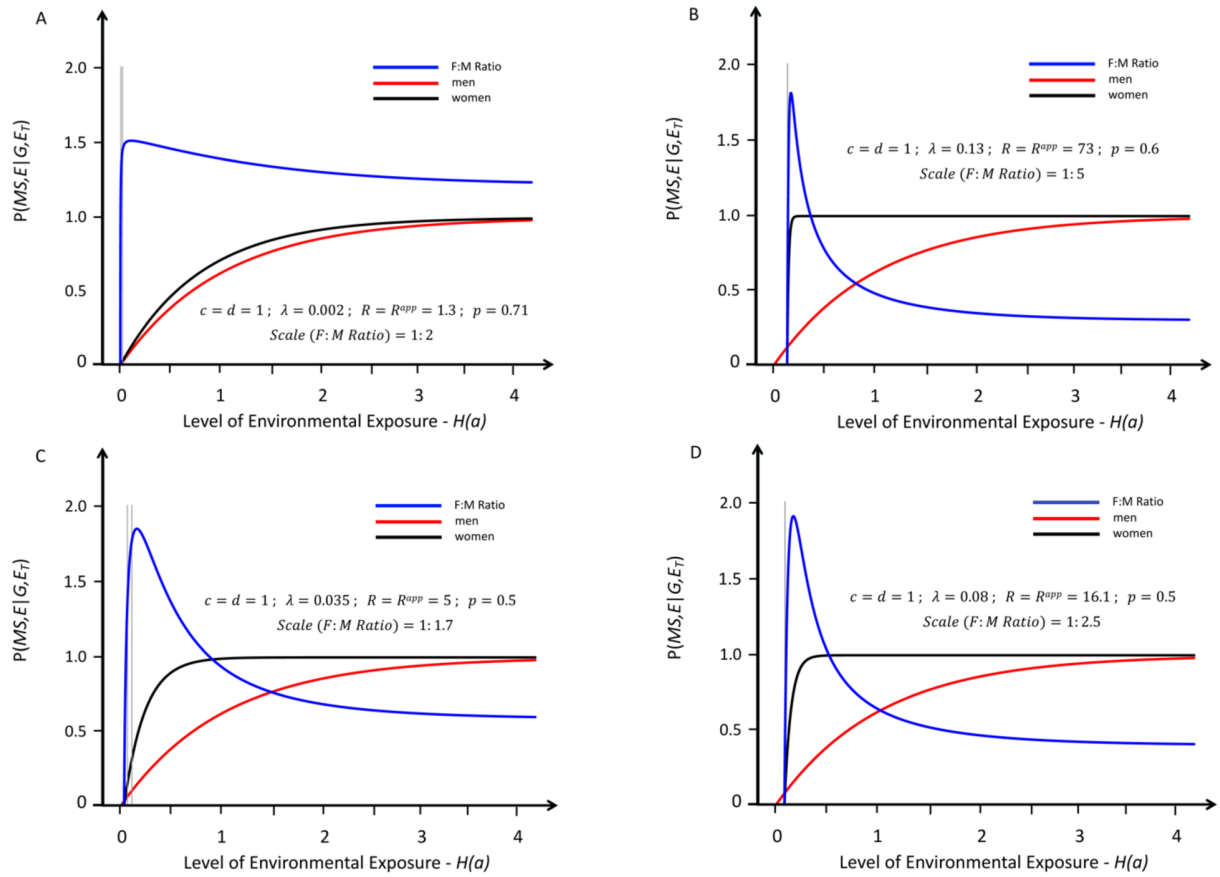

Figure 3. Response curves for the likelihood of developing MS in genetically susceptible *women* (black lines) and *men* (red lines) with an increasing probability of a "sufficient" environmental exposure – *see Methods #1B*. Like *Figure 1*, the curves depicted are also proportional ($R = R^{app}$), but, for these, the environmental threshold in *women* is greater than that it is in *men* – i.e., these are conditions in which: ($\lambda > 0$). Also, all these response curves represent actual solutions and reflect conditions in which ($\boldsymbol{c} = \boldsymbol{d} = 1$) and, as discussed in *Methods #4C*, are representative of all conditions in which $\boldsymbol{c} = \boldsymbol{d} < 1$). Moreover, with increasing values from ($R^{app} \geq 1.3$), which is the minimum value of (R^{app}) for any solution – which is depicted in *Figure A*. The blue lines represent the change in the (*F:M*) *sex ratio* (plotted at various scales) with increasing exposure. The thin grey vertical lines represent the portion of the response curve (for the depicted solution), which represents the actual change in the (*F:M*) *sex ratio* that occurred between *Time Periods #1 & #2*). To account for the observed increase in the (*F:M*) *sex ratio*, these curves require the Canadian observations [6] to have been made over a very small portion the response curve – i.e., for most of these response curve, the (*F:M*) *sex ratio* is decreasing. Also, for each of these response curves, including the maximum difference in the environmental threshold (i.e., $\lambda \leq 0.13$) under conditions of ($\boldsymbol{c} = \boldsymbol{d} = 1$), which is depicted in *Figure B*, the ascending portion of the curve (which reflects and increasing *F:M sex ratio*) is very steep – a circumstance indicating that the portion of the response curve available for fitting the Canadian data [6] is quite narrow.

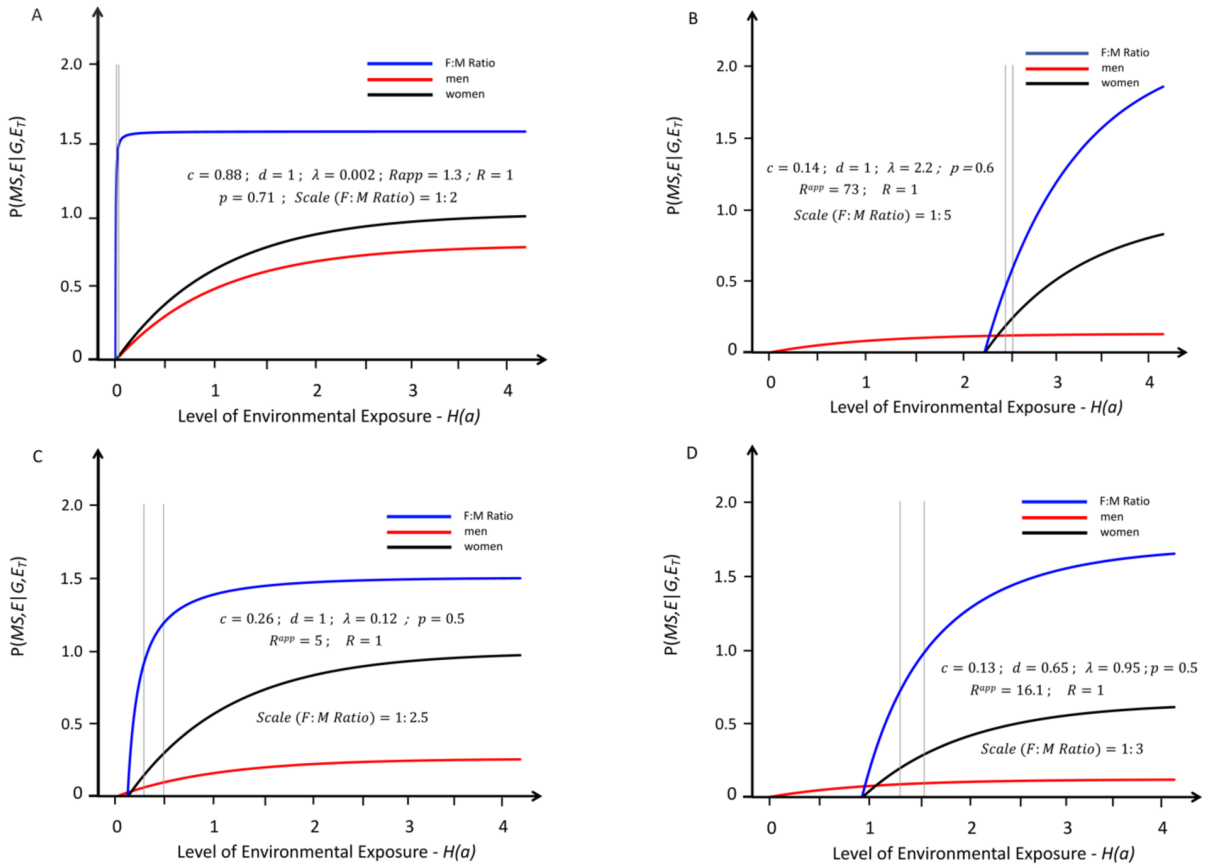

Figure 4. Response curves for the likelihood of developing MS in genetically susceptible *women* (black lines) and *men* (red lines) with an increasing probability of a "sufficient" environmental exposure – *see Methods #1B*. Like *Figure 1*, the curves depicted are also proportional ($R = 1$), but, for these, the environmental threshold in *women* is greater than that it is in *men* – i.e., these are conditions in which: ($\lambda > 0$). Also, these curves represent the same solutions as those depicted in *Figure 3* except that these are for conditions in which ($\boldsymbol{c} < \boldsymbol{d} \leq 1$). The blue lines represent the change in the (*F:M*) *sex ratio* (plotted at various scales) with increasing exposure. The thin grey vertical lines represent the portion of the response curve (for the depicted solution), which represents the actual change in the (*F:M*) *sex ratio* that occurred between *Time Periods #1 & #2*). Unlike the curves presented in *Figure 3*, however, an increase in the (*F:M*) *sex ratio* with increasing exposure is observed for any two-point interval along the entire response curves and, except for *Figure A*, the grey lines are clearly separated.

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. (2021) The nature of genetic and environmental susceptibility to multiple sclerosis. *PLoS One* 16(3): e0246157.

4. Ebers GC, Sadovnick AD, Risch NJ, the Canadian Collaborative Study Group. (1995) A genetic basis for familial aggregation in multiple sclerosis. *Nature* 377:150-151.

5. Sadovnick AD, Risch NJ, Ebers GC. (1998) Canadian collaborative project on genetic susceptibility to MS, phase 2: Rationale and method. Canadian Collaborative Study Group. *Can J Neurol Sci*. 25:216-21.

6. 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.

7. 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.

8. 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.

9. 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.

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

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. Witte JS, Carlin JB, Hopper JL. (1999) Likelihood-based approach to estimating twin concordance for dichotomous traits. *Genetic Epidemiol*. 16:290–304.

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

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

19. Machin G. (2009) Familial monozygotic twinning: A report of seven pedigrees. *Am J Med Genet* 151C:152-154.

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

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

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

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

24. 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.

25. 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.

26. 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

27. 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.

28. 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.

29. Walton C, King R, Rechtman L, et al. (2020) Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, 3rd Edition. *Mult Scler* 26:1816-1821.

30. Koch-Henriksen N, Sørensen PS. (2010) The changing demographic pattern of multiple sclerosis epidemiology *Lancet Neurol*. 9:520-532.

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

32. 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.

33. 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.

34. 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.

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

36. 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.

37. 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.

38. Jacobson HI. (1969) The maximum variance of restricted unimodal distributions. *Ann Math Stat* 40:1746–52.

39. Fisher LD, van Belle G. (1993) Biostatistics: A Methodology for the Health Sciences, John Wiley & Sons, New York, pp. 369-373 & 786-829.

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

41. Willer CJ, Dyment DA, Sadovnick AD, et al. (2005) Timing of birth and risk of multiple sclerosis: population based study. *Br Med J* 330:120-124.

42. Staples J, Ponsonby AL, Lim L. (2010) Low maternal exposure to ultraviolet radiation in pregnancy, month of birth, and risk of multiple sclerosis in offspring: longitudinal analysis. *Br Med J* 340:c1640.

43. Fiddes B, Wason J, Kemppinen A, Ban M, Compston A, Sawcer S. (2013) Confounding underlies the apparent month of birth effect in multiple sclerosis. *Ann Neurol.* 73:714-717.

44. Templer DI, Trent NH, Spencer DA, et al. (1992) Season of birth in multiple sclerosis. *Acta Neurol Scand* 1992;85:107-109.

45. Pantavou KG, Bagos PG. (2020) Season of birth and multiple sclerosis: a systematic review and multivariate meta-analysis *J Neurol.* 267:2815-2822.

46. Kurtzke JF, Beebe GW, Norman JE (1979) Epidemiology of multiple sclerosis in U.S. veterans: 1. Race, sex, and geographic distribution. *Neurology* 29:1228-1235

47. Sabel CE, Pearson JF, Mason DF, et al. (2021) The latitude gradient for multiple sclerosis prevalence is established in the early life course. *Brain* 144:2038-2046.

48. Bjornevik K, Cortese M, Healy BC, et al. (2022) Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science* 375:296-301.

49. Winter JR, Taylor GS, Thomas OG, Jackson C, Lewis JEA, Stagg HR. (2019) Predictors of Epstein-Barr virus serostatus in young people in England. *BMC Infect Dis* 19:1007(pp.1-9)

50. Munger KL, Zhang SM, O'Reilly E et al. (2004) Vitamin D intake and incidence of multiple sclerosis. *Neurology* 62:60–65.

51. Aujla RS, Allen PE, Ribbans WJ. (2019) Vitamin D levels in 577 consecutive elective foot & ankle surgery patients. *Foot Ankle Surg*:310-315.

52. Vallejo MS, Blümel JE, Arteaga E, et al. (2020) Gender differences in the prevalence of vitamin D deficiency in a southern Latin American country: a pilot study. *Climacteric* 23:410-416.

53. Sowah D, Fan X, Dennett L, Hagtvedt R, Straube S. (2017) Vitamin D levels and deficiency with different occupations: a systematic review. *BMC Public Health* 22;17:519.

54. Sundström P, Nyström L, Hallmans G. (2008) Smoke exposure increases the risk for multiple sclerosis *Eur J Neurol*. 15:579-583.

55. 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.

56. 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.

57. 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.

58. 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.

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

60. 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.

61. Goodin DS, Khankhanian P, Gourraud PA, Vince N. (2021) Genetic susceptibility to multiple sclerosis: interactions between conserved extended haplotypes of the MHC and other susceptibility regions. *BMC Med Genomics* 14:183

62. Kuri A, Jacobs BM, Vickaryous N, Pakpoor J, Middeldorp J, Giovannoni G, Dobson R. (2020) Epidemiology of Epstein-Barr virus infection and infectious mononucleosis in the United Kingdom *BMC Public Health*. 20:912 (pp.1-9)

63. Munger KL, Levin LI, Hollis BW, Howard NS, Ascherio A. (2006) Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. *JAMA*. 296:2832-2838

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

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

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

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

68. 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.

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

70. Green B. (2020) Until the end of time. Alfred A Knopf, Penguin Random House, New York, USA

71. Layzer D. (2021) Why we are free: Consciousness, free will and creativity in a unified scientific worldview. Information Publisher. ISBN-10 0983580251.

72. Chambers J, Mitton J. (2014) From dust to life: The origin and evolution of our solar system. Princeton, NJ, USA: Princeton University Press. pp. 96–107.

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

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

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

76. 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

77. Gragert 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.

78. Goodin DS, Oksenberg JR, Douillard V, Gourraud PA, Vince N. (2021) Genetic susceptibility to multiple sclerosis in African Americans. *PLoS One*. 16(8):e0254945.

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

Parameter	Definition
(Z)	Set of all (N) individuals (i.e., unique genotypes) in the population
$(G), (G^c)$	Subsets of individuals (genotypes) in (Z) who have <u>*any*</u> non-zero chance (G) or no chance (G^c) of developing MS
$(G_{w^s}), (G_{dw^s})$	Subset of susceptible *women*: $(G_{w^s}) = (F, G)$; (G_{dw^s}) = the dth susceptible *woman* in (G)
$(G_i), (G_{is}), (G_{ia})$	Genotypes or subsets either of the ith, or an *i-type*, individual in (G): (G_i) = full unique genotype; (G_{is}) = genotype of all *MS*-related factors (G_{ia}) = genotype of only autosomal *MS*-related factors
$\{G_s\}, \{G_a\}$	Families of sets: $-$ $\{G_s\}$ includes all subsets (G_{is}) within (G); and $\{G_a\}$ includes all subsets (G_{ia}) within (G)
$(M), (F)$	Subsets of *men* (M) and *women* (F) in (Z) -- $P(M) + P(F) = 1$
$(G1), (G2)$ †	"*High*" $(G1)$ and "*low*" $(G2)$ penetrance" subsets in (G) – *see Text*
$(G1'), (G2')$ †	"*High*" $(G1')$ and "*low*" $(G2')$ penetrance" subsets in (F, G) or (M, G)
$(MZ), (DZ), (S)$	Subsets *MZ*-twins, *DZ*-twins, or non-twin siblings (*S*) in (Z)
(E_T) †	The prevailing environmental conditions during the *Time Period* (T)
$\{E_i\}$ †	Family of sets of environmental exposures, each of which is sufficient, by itself, to cause MS in the ith individual in (G) – *see Text*
$\{E_{iw}\}$ †	Family of sets of more "*intense*" environmental exposures, within $\{E_i\}$, required by an *i-type woman* – $(E_{iw}) \subset (E_i)$ – *see Text*
(E) †	Union of all disjoint events $(\{E_i\}, G_i)$ – *see Text*
(E_w) †	Union of all disjoint events $(\{E_{iw}\}, G_i)$ – $(E_w) \subset (E)$ – *see Text*
$(E_{pop}), (E_{sib}), (E_{twn})$	Distinct parts of a sufficient environmental exposure, equally likely to be shared by anyone (E_{pop}), more or less likely to be shared by siblings (E_{sib}) and more or less likely to be shared by twins (E_{twn}) – *see Text*
(MS) †	Subset of individuals in (Z) who either have, or will develop, *MS*
(MZ_{MS}), (DZ_{MS}), (S_{MS})	Subsets of *MZ* co-twins (MZ_{MS}), *DZ* co-twins (MZ_{MS}), or non-twin co-siblings (S_{MS}) who either have, or will develop, *MS* – *see Text*
$P(MS)$ †	Life-time probability of developing *MS* for a member of (Z):
$P(MS \mid MZ_{MS})$ † $P(MS \mid DZ_{MS})$ † $P(MS \mid S_{MS})$ †	$P(MS)$ for a proband, randomly selected from (Z), who has a co-twin or co-sibling in the (MZ_{MS}), (DZ_{MS}), or (S_{MS}) subsets – *see Text*
$P(MS \mid IG_{MS})$ †	$P(MS \mid MZ_{MS})$ adjusted for the shared environment of *MZ*-twins
$P(MZ_{MS})$ † $P(IG_{MS})$ †	$P(MS)$ for the co-twin from an *MZ* twin-pair, without considering the proband's circumstances. $P(MZ_{MS}) = P(IG_{MS}) = P(MS)$ – *See Text*

† Parameters that vary with environmental conditions (E_T), which is indicated in different manners in the *Text*, depending upon context. For example, during: (E_T) = *Time Period #2*:

$$P(MS) = P(MS \mid Z, E_T) = P(MS \mid E_T) = P(MS \mid Z)_2 = P(MS)_2$$

The subscript "MS" indicates that the non-proband relative preceding the subscript has or will develop MS

Table 2. Principal Parameter Abbreviations†

Parameter	Definition
(F, w) & (M, m)	Alternate indicators for: female/women (F, w) and: male/men (M, m)
$Subscripts$ (1,2)	Indicators for either *Time Period #1* or the "current" *Time Period #2* -- except for the parameters: $x_1, x_2, x_1', x_2', y_1', and\ y_2'$ -- *see below*
x	Penetrance of MS for the (G) subset -- $x = P(MS \mid G)$
x_i	Penetrance of the (i^{th}) genotype in (G) -- $\forall G_i \in G$: $P(MS \mid G_i) = x_i$
X	Set of all individual penetrance values in (G) -- $X = \{x_i\}$
x'	Adjusted *MZ*-twin concordance for MS -- $x' = P(MS \mid IG_{MS})$
σ_X^2	Variance of the set (X) -- $Var(X)$
x_1	Penetrance of MS for $(G1)$ subset -- $x_1 = P(MS \mid G1)$
x_2	Penetrance of MS for $(G2)$ subset -- $x_2 = P(MS \mid G2)$
Zw	Failure probability in *women*: $Zw = P(MS, E_w \mid F, G) = x_1$
Zm	Failure probability in *men*: $Zm = P(MS, E \mid M, G) = x_2$
x_1'	Adjusted *MZ*-twin concordance for $(G1)$ subset -- $x_1' = P(MS \mid G1, IG_{MS})$
x_2'	Adjusted *MZ*-twin concordance for $(G2)$ subset – $x_2' = P(MS \mid G2, IG_{MS})$
y_1'	*MZ*-twin concordance for the $(G1)$ subset -- $y_1' = P(MS \mid G1, MZ_{MS})$
y_2'	*MZ*-twin concordance for the $(G2)$ subset -- $y_2' = P(MS \mid G2, MZ_{MS})$
u	Odds of a sufficient environmental exposure -- $u = P(E) / [1 - P(E)]$
q_m^{min}, q_w^{min}	*"Minimum"* exposure level change in *men* (q_m^{min}) and *women* (q_w^{min})
q_m , q_w	*"Actual"* exposure level change in *men* (q_m) and *women* (q_w)
$h(u)$, $k(u)$	Hazard functions for *men* – $h(u)$; and for *women* – $k(u)$
R	*"Actual"* proportionality factor (*if proportional*) -- $k(u) = R * h(u)$
R_i , R_{dw^s}	*"Actual" value of R* in *i-type* (R_i) and *individual* (R_{dw^s}) *women*
R^{app}	"*Apparent*" value of R -- $R^{app} = q_w^{min} / q_m^{min}$
λ_w , λ_m	Environmental threshold to develop MS in *women* (λ_w) and *men* (λ_m)
λ , λ_i	Threshold differences, generally: $\lambda = (\lambda_w - \lambda_m)$; and for *i-types* ($\lambda_i$)
C	Ratio of *P*(*MS*) at *Time #1* to that at *Time #2* -- $C = P(MS)_1 / P(MS)_2$
C_F , C_M	The ratio (C) for *women* (C_F) ; and for *men* (C_M)
p , p'	Proportion of *women* in subset $(G) - p$; and in subset $(MS, G) - p'$
r , s	Enrichment of genotypes – *women*: $r = (x_1' / x_1)$; *men*: $s = (x_2' / x_2)$
$\boldsymbol{c}$, $\boldsymbol{d}$	Limits for the exponential response curves in *men* ($\boldsymbol{c}$) and *women* ($\boldsymbol{d}$)
s_a	Factor to adjust $P(MS \mid MZ_{MS})$ because of the shared twin environment $P(MS \mid IG_{MS}) = P(MS \mid MZ_{MS}) / s_a$; $s_a \geq 1$ – *see Text*

† Each of these parameters (except $h(u), k(u), R, R_i, R_{w_j}, \lambda, \lambda_w, \lambda_m, p, \boldsymbol{c}, \& \boldsymbol{d}$) vary with different environmental conditions (E_T), which is indicated in various ways in the *Text– see Table 1*.

Table 3. MS Associations for Class I and Class II *HLA*-Haplotypes in *Men* and *Women*[*]

Haplotype	***OR*[†] *(Women)***	***p***	***OR*[†] *(Men)***	***p***
*DRB1*15:01~DQB1*06:02 – 1 copy* [‡]	3.1	< E-182	2.7	< E-82
*DRB1*15:01~DQB1*06:02 – 2 copies* [‡]	6.5	< E-110	6.2	< E-70
*DRB1*03:01~ DQB1*02:01 – 1 copy* [‡]	1.1	< 0.05	1.2	< 0.01
*DRB1*03:01~ DQB1*02:01 – 2 copies* [‡]	2.9	< E-21	2.2	< E-7
*A*02:01~C*05:01~B*44:02 – 1 copy*	0.6	< E-7	0.6	< E-3

* Class I or Class II haplotypes in the WTCCC within the human leukocyte antigen (*HLA*) region on the short arm of Chromosome 6 [60,61]. These haplotypes include either *HLA* Class I (*A, C & B*) or *HLA* Class II (*DRB1 & DQB1*) alleles.

† Odds ratio (*OR*) for MS in *men* and *women* having either 1 or 2 copies of each haplotype except for the *A*02:01~C*05:01~B*44:02* haplotype, which had too few observations of the *2-copy* data. Each was compared to individuals having no copies of the other 2 *haplotypes* (95% *CI* range in parenthesis). The *p-values* are expressed in scientific notation as powers of 10 (E).

‡ The difference in *OR* between possessing *1* or *2 copies* of the *DRB1*15:01~DQB1*06:02 01* haplotype was significant for both *women* ($p < 10^{-13}$) and *men* ($p < 10^{-10}$). Similarly, the difference in *OR* between possessing *1* or *2 copies* of the *DRB1*03:01~DQB1*02:01* haplotype was significant for both *women* ($p < 10^{-14}$) and *men* ($p < 0.001$).

Table 4. MS Associations for Extended *HLA*-Haplotypes Region in *Men* and *Women*

CEH Name*	***CEHs***** ***A~C~B~DRB1~DQB1~SNP***	***OR*† *(Women)***	***p***	***OR*† *(Men)***	***p***
c1	*01:01~07:01~08:01~03:01~02:01~a6*	3.5	< E-7	1.8	0.02
c2‡	*03:01~07:02~07:02~15:01~06:02~a1*	2.7	< E-78	2.5	< E-36
c3‡	*02:01~07:02~07:02~15:01~06:02~a1*	1.9	< E-17	1.9	< E-9
c5	*02:01~05:01~44:02~04:01~03:01~a3*	0.5	< E-9	0.5	< E-6

* Arbitrary names for the four most common MS-associated conserved extended haplotypes (*CEHs*) in the WTCCC within the human leukocyte antigen (*HLA*) region on the short arm of Chromosome 6 [61].

** These *CEHs* include both *HLA* Class I (*A, C & B*) and *HLA* Class II (*DRB1 & DQB1*) alleles in addition to the haplotypes of 11 single nucleotide polymorphisms (*SNPs*) spanning the *HLA* Class II chromosomal region [60].

† Odds ratio (*OR*) for MS in *men* and *women* having 2 copies of the (*c1*) *CEH* (only homozygotes were *MS*-associated) or any number of copies of the (*c2*), (*c3*), or (*c5*) *CEHs* (both homozygotes and heterozygotes were *MS*-associated) compared to having no copies of the other three *CEHs* (95% *CI* range in parenthesis). The *p-values* are expressed in scientific notation as powers of 10 (E).

‡ The difference in *OR* between the (*c2*) and (*c3*) *CEHs* was significant for both *women* (*p=0.0001*) and *men* (*p=0.03*).