UNIVERSITÀ DEGLI STUDI DI PAVIA

Department of Brain and Behavioral Sciences

Multiple Sclerosis Heritability Estimation on Sardinian Ascertained Extended Families Using Bayesian Liability Threshold Model

PhD Candidate: Andrea Nova

Tutor: Prof. Luisa Bernardinelli

PhD in Psychology, Neuroscience and Data Science

Curriculum: Data Science in Biomedicine

XXXVI Cycle, A.A. 2022-2023

ABSTRACT

INTRODUCTION: Narrow-sense heritability (h^2) measures the proportion of phenotypic variability observed in a specific population that is attributable to the sum of additive genetic effects. Heritability studies represent an important tool to investigate the main sources of variability for complex diseases, whose etiology involves both genetics and environmental factors.

AIM: The present work aimed to estimate multiple sclerosis (MS) narrow-sense heritability, on a liability scale, using 24 extended families ascertained from affected probands sampled in the Sardinian province of Nuoro, Italy. The sources of MS liability variability were also investigated among shared environmental effects, sex, and categorized year of birth (<1946, ≥1946). The latter can be considered a proxy for different early environmental exposures.

METHODS: A Bayesian liability threshold model (Bayesian-LTMH) was developed to estimate heritability for binary phenotypes making use of ascertained family-based samples, overcoming the limitations of the previously suggested EM algorithm. The Bayesian approach allows one to obtain the posterior distribution and credibility interval (CI) for heritability adjusted for potential confounders, such as shared environmental effects. The performance of Bayesian-LTMH was evaluated via simulation experiments and was then implemented to analyze the Sardinian families to obtain posterior distributions for the parameters of interest adjusting for ascertainment bias.

RESULTS: Simulation studies highlighted the accuracy and precision of Bayesian-LTMH, other than the dramatic improvement in computational efficiency compared to the approach based on the EM algorithm. The analysis of the Sardinian sample highlighted categorized year of birth as the main explanatory factor, explaining ~70% of MS liability variability (median value = 0.69, 95% CI: 0.64, 0.73), while h^2 resulted near to 0% (median value = 0.03, 95% CI: 0.00, 0.09). By performing a year of birth-stratified analysis, a high h^2 was found only in individuals born on/after 1946 (median value = 0.82, 95% CI: 0.68, 0.93), meaning that the genetic variability had a high explanatory role only when focusing on this subpopulation.

CONCLUSIONS: Overall, the results obtained highlighted early environmental exposures, in the Sardinian population, as a meaningful factor involved in MS to be further investigated. These environmental factors are likely linked to the westernization process that occurred in Sardinia after World War II. Among these the malaria eradication program has been previously pinpointed, under the light of the hygiene hypothesis, as a key factor to explain the dramatic rise in MS incidence in the last decades.

CONTENTS

LIST OF FIGURES AND TABLES

FIGURES

TABLES

Table 1. Descriptive statistics for the sampled posterior distributions obtained fitting Bayesian liability threshold model on the 200 simulated datasets within each different scenario.

Table 2. Descriptive statistics for the 24 Sardinian extended families.

Table 3. Descriptive statistics for the 118 Multiple Sclerosis cases in the Sardinian families.

Table 4. Kinship relationships between the 118 multiple sclerosis cases.

Table 5. Posterior distributions summary statistics for parameters included in the Bayesian-LTMH applied to the Sardinian families.

Table 6. Posterior distributions summary statistics for parameters included in the Bayesian-LTMH applied to the Sardinian families stratified by year of birth on different environment conditions.

1. INTRODUCTION

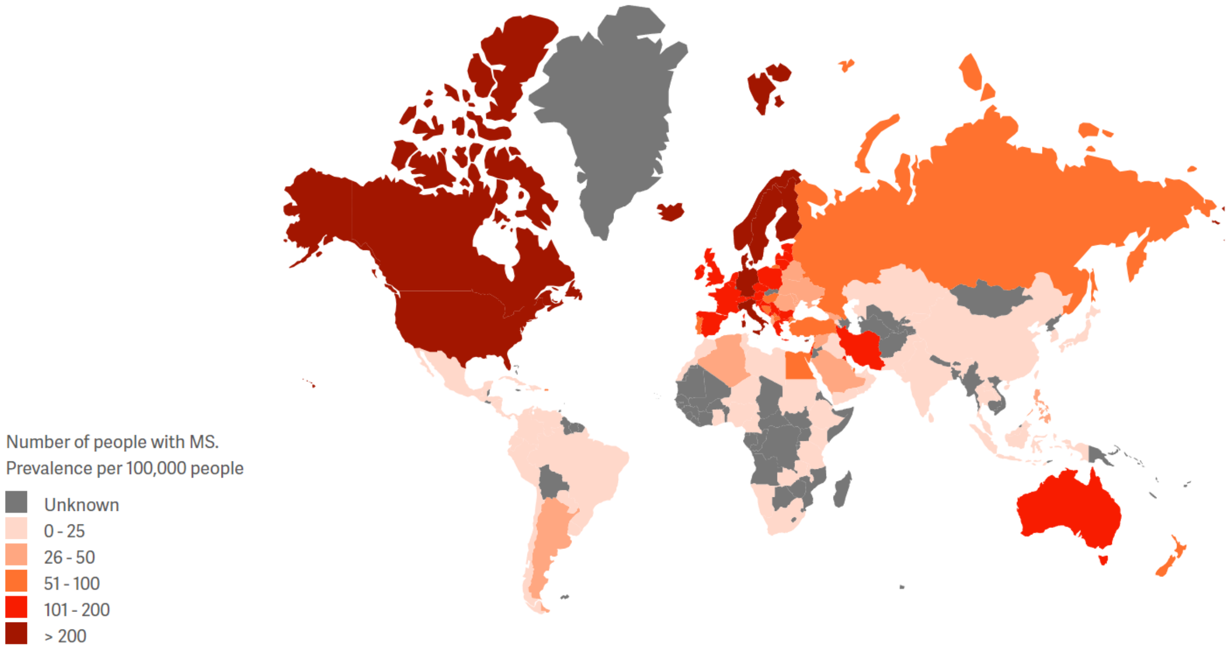
1.1 Background on Multiple Sclerosis

Affecting over 2.8 million people worldwide [1], Multiple Sclerosis (MS) is a multifactorial disease with progressive neurodegeneration characterized by chronic inflammation and demyelination in the central nervous system (CNS) [2,3].

1.1.1 Epidemiology

MS may onset at all ages of life, even before 18 years old, although initial symptoms typically present between 20 and 40 years of age. MS affects women approximately twice as often as men [4], and it's showing increasing incidence and prevalence worldwide over time [1,5]. Moreover, MS represents the most common cause of non-traumatic neurological disability in young adults, and life expectancy is reduced by 7 to 14 years compared with the general, healthy population.

The distribution of MS on a global scale exhibits a distinct pattern that correlates with geographical latitude. Prevalence rates of MS tend to be highest in regions located further away from the equator [6,7]. MS prevalence can vary significantly within specific regions or populations. For example, in some isolated communities there may be a lower incidence of MS, while in others, particularly indigenous populations, like the Sámi in Scandinavia or in certain areas of Sardinia, high prevalence rates are observed [8]. Sardinia, an island in the Mediterranean, exhibits a distinctive pattern of MS prevalence. It has a peak prevalence rate of around 300 cases per 100,000 individuals, which is among the highest in the world [9]. This differs from the prevalence observed in continental Italy [10]. As shown in **Figure 1,** there seems to be a correlation between MS prevalence and socioeconomic status, as, generally, MS tends to be less prevalent in Asian countries and becomes more common in regions and populations such as the USA, European countries, and Australia [8].



There are 2.9 million people living with MS worldwide.

Figure 1. The geography of Multiple Sclerosis: prevalence per 100,000 population in 2023.
Source: https://www.atlasofms.org/map/global/epidemiology/number-of-people-with-ms

1.1.2 Symptoms and disease course

The manifestation of MS is highly variable, as a wide range of symptoms can vary among individuals. Common symptoms include: fatigue, visual impairment, numbness and tingling, muscle weakness, balance and coordination issues, pain, cognitive and emotional changes, bladder and bowel dysfunction, tremors, and sexual dysfunction [11,12]. These clinical symptoms are useful to clinicians to diagnose MS, in conjunction with other neurological exams and diagnostic tests such as magnetic resonance imaging (MRI) of the brain and spinal cord. Moreover, MS doesn't always progress through distinct stages in a linear fashion. However, it is often described in terms of the following generalized stages or clinical courses [12,13]:

- Relapsing-Remitting MS (RRMS): in most individuals with MS, the disease presents itself as episodes of neurological dysfunction that spontaneously improve. Between relapses, individuals may experience stable periods.
- Secondary Progressive MS (SPMS): Some individuals with RRMS eventually transition to SPMS. In this stage, there is a gradual worsening of symptoms and disability, with or without relapses and remissions. The disease becomes more steadily progressive.
- Primary Progressive MS (PPMS): This form of MS is characterized by a gradual and steady progression of disability from the onset, without distinct relapses and remissions. It is less common than RRMS.

Not all individuals with MS will experience the same course or stages. Some may remain in the RRMS stage throughout their lives, while others may progress to SPMS or PPMS.

Currently, there are 15 FDA-approved medications for RRMS, of which 14 are also used against SPMS and only one against PPMS [14]. This underscores the need for better therapeutic options for progressive forms of the disease [15]. Additionally, existing treatments that target inflammation, modulate the immune system, or suppress immune responses tend to show effectiveness in the initial phases of the disease. However, their benefits become limited once patients progress into the later stages. While the increasing range of treatments aimed at diminishing disability and prolonging the lives of individuals with MS, a definitive cure is yet to be discovered, and our understanding of the disease's underlying causes and etiology remains incomplete and not fully understood.

1.1.3 MS pathogenesis

A defining feature of MS pathology is the presence of localized areas known as plaques where demyelination occurs. These plaques are closely associated with episodes of MS relapses, during which inflammation leads to the removal of myelin in both white and gray matter of the central nervous system. This demyelination process involves the infiltration of macrophages, as well as T and B lymphocytes [16]. Remyelination serves as the natural restorative mechanism for reversing demyelination, and it's hypothesized that axons that have been remyelinated are shielded from degeneration. Unfortunately, remyelination varies significantly among MS patients, and, for reasons not yet fully understood, it often either fails or remains incomplete [17]. In fact, more than two-thirds of patients eventually progress to SPMS. This phase of the disease is thought to be primarily driven by neurodegeneration, leading to a slow and irreversible accumulation of disability, particularly affecting a person's ability to walk and their cognitive function. As previously mentioned, in a small percentage of MS patients, this progressive stage begins right from the onset of the disease, i.e., PPMS. The underlying mechanisms behind primary and secondary progression are not yet fully understood [18]; however, various lines of evidence suggest that the symptoms of progressive MS can vary depending on the location of neurological lesions. The dysregulation of the blood-brain barrier (BBB), and the activation of myelin-reactive T cells in the CNS periphery are among the earliest cerebrovascular abnormalities observed in the brains of individuals with MS. These events indicates the beginning of the inflammation, leading to infiltration of inflammatory cells through the BBB, along with the migration of activated white blood cells through the endothelial cells of blood vessels with consequent release of inflammatory cytokines [17]. The observed changes in the arrangement of junctional proteins on the BBB, such as cellular adhesion molecules (CAMs), are widely recognized to occur during neuroinflammatory and infectious events [19]. Consequently, the presumed underlying mechanism of BBB disruption is associated with autoimmune reactions, particularly in individuals who have a genetic predisposition to such reactions. In MS, the disruption of the BBB is believed to be temporary, and the subsequent evolution and formation of lesions occur intermittently. This process includes additional episodes of BBB leakage, immune-mediated demyelination, and varying degrees of axonal damage. (**Figure 2**).

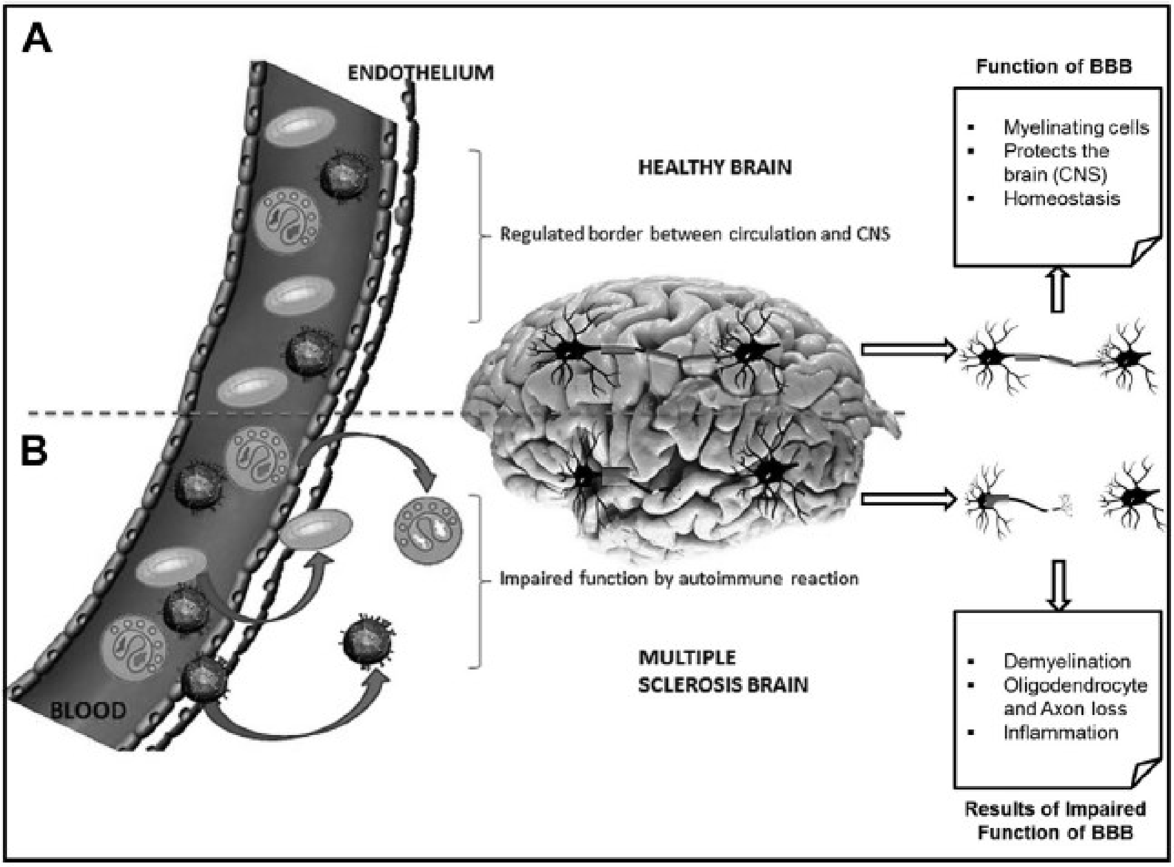

A
ENDOTHELIUM
HEALTHY BRAIN
Regulated border between circulation and CNS
Function of BBB
▪ Myelinating cells
▪ Protects the brain (CNS)
▪ Homeostasis
B
Impaired function by autoimmune reaction
MULTIPLE SCLEROSIS BRAIN
BLOOD
▪ Demyelination
▪ Oligodendrocyte and Axon loss
▪ Inflammation
Results of Impaired Function of BBB

Figure 2. Blood-brain-barrier (BBB) comparison in healthy and Multiple Sclerosis brains. (A) BBB is an extremely specific endothelial structure, which functions as a protective mediator of the brain, separating the circulating blood components from neurons, keeping the homeostasis and functional myelination cells. (B) In pathological conditions, BBB dysfunction leads to immune cell infiltration followed by larger inflammatory responses. In the brain, lesions of Multiple Sclerosis patients include demyelination, axonal loss, and neurodegenerative process.

The inflammatory process in the CNS primarily affects oligodendrocytes, which are responsible for myelinating CNS cells, as well as neurons. Early in the disease progression, this inflammatory state then leads to characteristic neuronal death, demyelination and axonal loss and becomes the predominant characteristic as the disease advances [17]. The prevailing hypothesis suggests that this axonal loss is the primary mechanism responsible for the progressive disability observed in MS [8,20] (**Figure 3**).

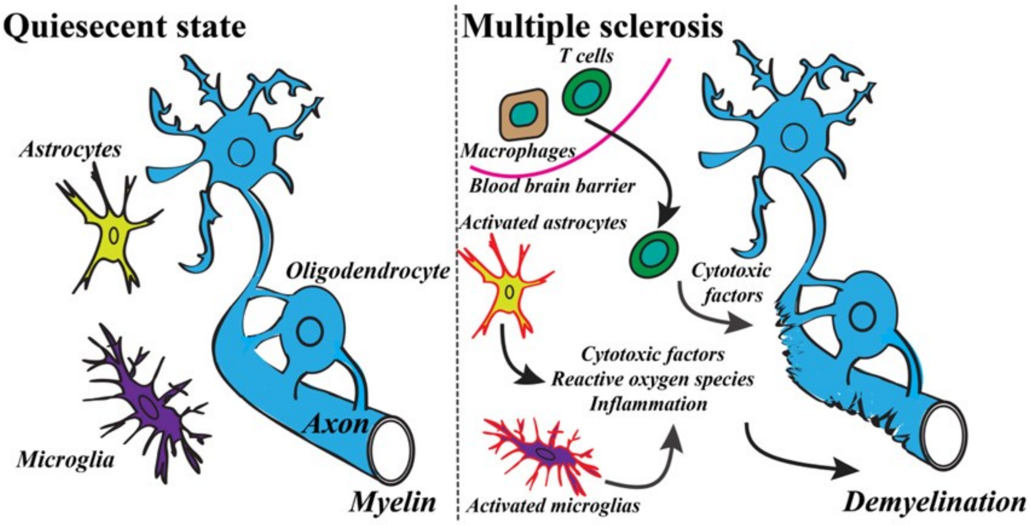


Figure 3. Basic mechanism in the development of Multiple Sclerosis, which includes a variety of inflammatory responses and activation of specific cell types.

Moreover, during the inflammatory response in the CNS, antigen-presenting cells (APCs), which include B cells, microglia, macrophages, and dendritic cells, present potential autoantigens like myelin basic protein or myelin-oligodendrocyte protein through MHC class II molecules [21]. This presentation triggers further release of cytokines and chemokines fuelling continuous inflammation [11,22]. This process ultimately results in demyelination, edema, and distinct white matter lesions, which are typically found in areas such as the subcortical or periventricular white matter, optic nerve sheaths, brainstem, and spinal cord [15,17,23,24]. Hence, MS is thought to arise from a complex interaction involving BBB disruption, inflammatory cells such as microglia and astrocytes within the CNS, as well as autoreactive T cells from the peripheral immune system that migrate to the CNS [25]. This inflammation triggers damage to brain tissue, and this tissue damage subsequently exacerbates inflammation creating a harmful cycle. This flow is depicted in **Figure 4**.

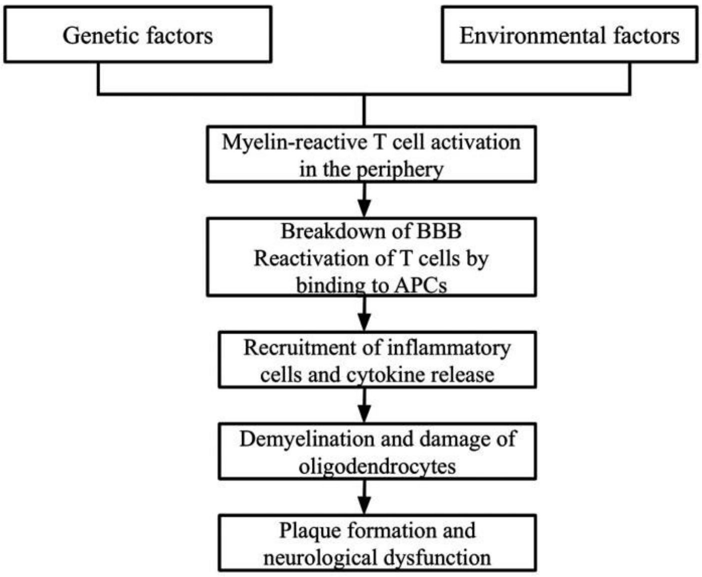

Genetic factors
Environmental factors
Myelin-reactive T cell activation in the periphery
Breakdown of BBB
Reactivation of T cells by binding to APCs
Recruitment of inflammatory cells and cytokine release
Demyelination and damage of oligodendrocytes
Plaque formation and neurological dysfunction

Figure 4. Diagram depicting the potential pathogenesis of Multiple Sclerosis. Image credits: Ramya Talanki Manjunatha

The prevailing view is that the inflammatory process, potentially of autoimmune origin, is the primary cause of tissue damage in MS [23]. However, the limited comprehension of the mechanisms driving disease progression is evident in the limited available treatment choices for MS. Furthermore, numerous therapeutic approaches that had demonstrated success in typical experimental models of T cell-mediated inflammatory demyelinating diseases had no impact or even exacerbated the condition when applied to MS patients [11]. These inconsistencies indicate that MS is a very intricate disease. However, in addition to the "outside-in" theory, where T cells are initially activated in the periphery and then infiltrate the CNS, some studies propose an alternative perspective. This "inside-out" hypothesis suggests that the disease's origin lies within the CNS itself [26]. According to this view, MS begins with the primary degeneration of oligodendrocytes and myelin within the CNS, leading to the release of antigenic myelin components. These components, in turn, provoke an autoimmune response against myelin constituents. Regarding other hypotheses, in 2013, Chaudhuri [27] raised a perspective challenging the prevailing view that MS is primarily an inflammatory disease. He argued that the belief in MS being an inflammatory condition is largely based on an artificial experimental model (EAE) that induced demyelination after sensitization to myelin basic protein. According to Chaudhuri, the human disease's pathogenesis centres around blood vessels, particularly post-capillary venules, and areas where the BBB formed by vascular endothelial cells and astrocytic foot processes is disrupted. The assumption that the inflammatory changes near blood vessels in MS signify an autoimmune process remains unproven. Despite extensive efforts, no specific antigen or antibody has been definitively identified as a candidate for the cell-mediated or humoral immunopathogenesis of MS. Chaudhuri's conclusion is that inflammatory changes are secondary to tissue injury, and the disruption of the BBB is likely the key pathogenic event in MS. In individuals at risk of developing MS, the process of demyelination and neuroaxonal injury is heavily influenced by the astroglial response to oxidative and metabolic stress resulting from a locally disrupted BBB in the CNS. Various triggers, including infection, systemic inflammation, stress, physical trauma, or electrical injuries to the CNS, could initiate this disruption.

In summary, the pathogenesis of MS is highly heterogeneous and much more remains to be understood. However, MS is likely characterized by inflammatory processes triggered by a combination of genetic factors and environmental factors. This can be summarized in the significant theory put forth by Poser in 1986 [28]. According to his view, four elements are required to develop MS: i) genetic predisposition, ii) an environmental factor, likely related to viral immune-mediated events, iii) disruption of BBB function and iv) the ability to form myelinoclastic plaques within the CNS. In the next paragraph, the main MS risk factors highlighted in the scientific literature will be described, distinguishing among environmental and genetic factors.

1.1.4 MS risk factors

In the latter half of the 19th century, the observation of families with MS history provided the initial insights into the genetic aspect of the disease. When compared to the lifetime risk of 0.2% in the general population, first-degree relatives had a sevenfold increased risk of MS (2.5% excess lifetime risk). This risk is even more pronounced in identical twins (monozygotic), with a 30% likelihood of developing the disease [29]. In contrast, spouses and adoptees exhibit a risk that is similar to that of the general population, or in the case of adoptees, their original nuclear families. This consistency suggests that the driving force behind MS family history is genetic sharing [8]. On the other hand, the fact that the relative risk does not reach 100% even in identical twins suggests that other factors beyond DNA sequence identity must concur to create the conditions that cause or allow the dysregulation of the immune response associated with MS [8,24]. In this context, many aetiological factors have been identified in having an association with MS including genetic susceptibility, smoking, exposure to the Epstein-Barr virus (EBV), low exposure to sunlight (presumed to be mediated through vitamin D insufficiency), diet, body mass index (BMI), and microbiome as well as epigenetic signatures (e.g., DNA methylation patterns, histone modifications, and non-coding RNAs) [8,18,30].

1.1.4.1 Environmental factors

Growing evidence suggests that environmental factors play a significant role in the onset and progression of MS. Migrant studies, which examine MS risk in individuals who move from one region to another, provide further confirmation of the impact of environmental factors. A systematic review of these studies revealed two consistent patterns: migrants relocating from regions with a high MS risk to areas with a lower risk tended to have a lower prevalence of MS, especially when the migration occurred before the age of 15. Conversely, migrants moving from regions with a lower MS risk to higher-risk areas tended to maintain the lower MS risk of their home country, with no clear effect related to the age at which migration occurred [31]. These findings emphasize the significant influence of environmental factors on MS risk and suggest that early-life exposures may be particularly important [8]. The potential role of the main environmental risk factors on MS pathogenesis is now briefly described.

Low sunlight exposure and Vitamin D

As previously mentioned, there is a noticeable variation in the prevalence of MS based on geographical latitude, with higher rates observed at higher latitudes [32,33]. The role of vitamin D in explaining this latitude-related gradient was initially proposed. In humans, the primary source of vitamin D is exposure to ultraviolet B (UVB) radiation from sunlight, which varies in intensity depending on latitude and season. It's worth noticing that during the winter months, the lower intensity of UVB radiation may not provide sufficient support for vitamin D synthesis in certain regions, assuming equal sun exposure on an equivalent skin area [8]. Furthermore, vitamin D exerts significant effects on the immune system, and its immune-modulating properties have been observed in various cell-culture experiments [34]. These observations suggest potential biological mechanisms through which vitamin D may influence the risk of developing MS [35]. For instance, vitamin D has been found to reduce the production of interleukins IL-2 and IL-17, interferon-γ (IFN-γ), and it attenuates the cytotoxic activity and proliferation of CD4+ and CD8+ T cells. It also hinders B cell proliferation, plasma cell differentiation, and immunoglobulin production [33]. Until recently, the evidence supporting the idea that higher levels of vitamin D are associated with favourable effects on MS risk and a reduction in MS activity primarily relied on observational studies. However, Mendelian Randomization (MR) analyses support the notion that higher vitamin D levels play a causal protective role in MS risk, suggesting a direct cause-and-effect relationship between vitamin D and MS risk [1,36]. Notably, individuals with genetically lower levels of vitamin D are strongly associated with an increased susceptibility to MS. Moreover, maternal vitamin D deficiency is also a predisposing risk factor for the development of MS in pregnant women and their offspring[37]. Vitamin D supplementation is then advised to prevent the risk of MS in the general population [33,38]. Researchers have undertaken clinical trials to assess whether Vitamin D supplementation can effectively slow the progression of MS. Nevertheless, a recent meta-analysis of a randomized, double-blind, placebo-controlled clinical trial, examining the use of vitamin D as an adjunct therapy for MS, indicated that vitamin D did not demonstrate any therapeutic benefit in terms of reducing disability or the rate of relapses [39]. These findings may suggest that while Vitamin D levels in the norm are crucial to prevent the onset of MS, they may not have a curative effect once MS has already developed.

Viruses

Many virus infections have been proposed to play a role in MS pathogenesis. The severity of the viral infection depends on many different factors. Possibly, the most important is the interplay between virus and host immune mechanisms which are influenced by genetics and may sometimes have an impact on the development of symptomatic disease as a response to infectious agents. The clinical heterogeneity of MS and the diversity of MS plaques in the CNS suggest that there might be more than one infectious agent in the pathogenesis of this disease. The largest body of evidence during the last few years has accumulated around Epstein-Barr virus (EBV) and human herpesvirus 6 (HHV-6). Other associated agents include varicella-zoster virus (VZV), and human endogenous retroviruses (HERVs) [8,30]. The most consistent findings in relation to past infection is with EBV [40]. Infection with this herpes virus is most often asymptomatic in childhood but in adolescence and adulthood it is commonly symptomatic, causing infectious mononucleosis, which can be severe [41]. Large-scale population-based studies on EBV antibodies have shown consistently higher seroprevalence in patients with MS as compared to controls. In 2022, a significant epidemiological study provided compelling evidence regarding the role of the EBV in the development of MS [42]. The study, conducted over more than two decades and involving over 10 million individuals in the US Army, aimed to identify those diagnosed with MS and analyse their serum samples for anti-EBV antibodies. Individuals who became EBV seropositive (showing evidence of EBV infection) had a 32-fold increased risk of developing MS compared to those who remained seronegative (lacking evidence of EBV infection). Therefore, this study provided strong evidence linking EBV infection as an early and essential factor in the development of MS. However, MS prevalence is relatively low compared to the widespread presence of the EBV. One potential explanation is that the biological effects of EBV may vary depending on its genomic variability, implying the presence of potential gene-environment interactions explaining this discrepancy.

Body Mass Index

Both childhood and adult obesity are potential risk factors for MS explored extensively in research [43–45]. A review of MR studies highlighted consistent causal effects between increasing BMI and MS onset [1]. The connection between obesity and MS remains unclear, but various theories exist. Obesity involves persistent low-grade inflammation, with metabolic and immune cells interacting, possibly influencing MS risk. Childhood and adolescent obesity show proinflammatory markers, potentially impacting MS development. Adipokines like leptin, adiponectin, and resistin, along with gut microbiota and their role in immune responses, are suggested factors. Moreover, higher BMI levels could affect vitamin D levels and consequently MS risk, even though a significant causal effect was found even when adjusting for Vitamin D levels. Yet, the exact mechanism linking obesity and MS, whether through vitamin D or other pathways, is uncertain.

Smoking

While it is well-established that tobacco, which contains a high proportion of free radicals, can induce oxidative stress, and is implicated in numerous neurodegenerative disorders and autoimmune diseases, the impact of smoking on the immune system remains uncertain [46]. Cigarette components encompass pro-inflammatory effects, direct harm to tissues, and heightened apoptosis. Experiments on rats have shown that nicotine directly affects small parenchymal micro-vessels and tight junction proteins within the BBB, leading to increased permeable solute influx and alterations in blood flow to deep brain structures [47]. Additionally, cigarette smoke contains elevated concentrations of free radicals, including hydrogen cyanide, nitric oxide (NO), and carbon monoxide (CO), all of which contribute to oxidative damage in neural tissue [48]. Among a large North American cohort, it was observed that over 50% of MS patients were either current or former smokers. Furthermore, MS patients who smoked tended to be heavier smokers compared to the general population or individuals with other chronic conditions, and they often continued smoking after receiving a diagnosis [8]. However, while a meta-analysis of 14 case-control studies suggested an increased susceptibility to MS among smokers [49], this substantial risk factor did not seem to emerge from MR studies [1]. This contrast potentially hints at the presence of unmeasured confounding variables. The initiation of smoking, lifetime smoking, and smoking intensity all resulted in non-significant causal estimates in MR analyses, raising questions about the significance of smoking in the onset of MS.

Intestinal microbiota and diet

Low exposure to pathogens in early life has been suggested as a potential risk factor for the development of MS. This relates to the immune system's education, which can be influenced by viruses, parasites, and pathogenic bacteria, providing protection against autoimmunity. Likewise, dysbiosis of gut microbiota, characterized by specific microbial changes, has been observed in MS patients [37,50,51]. Changes in gut microbiota mainly derive from diet (particularly the types of fibres, fats, and sugars), infections, age, lifestyle factors like stress, sleep patterns, physical activity, and alcohol consumption, and finally genetics. These changes can lead to increased permeability of the intestinal and BBB, worsening MS severity. Moreover, short-chain fatty acids, which are produced during the fermentation of dietary fibre, have been implicated in MS pathogenesis, as these has been found to promote the differentiation of naïve CD4+ T cells into regulatory T cells, offering a beneficial effect in controlling MS symptoms. Instead, studies indicate that excessive consumption of saturated fats from animal sources may influence the risk of developing MS. Interestingly, regions with higher MS prevalence often have diets rich in gluten and milk[52].

1.1.4.2 Genetic Factors

The research on genetic factors associated with MS has primarily focused on the analysis of DNA sequences and the genetic differences among individuals, which then remain constant throughout their lives. However, in the past decade, a significant amount of information has emerged from the study of the epigenome, which involves heritable alterations in gene expression through modifications to the structure of DNA, without changing the underlying DNA sequence. Contrary to DNA sequence, the epigenome can change throughout the lifetime. These two aspects will be discussed separately.

Genome

The overall MS risk appears to be the result of the contributions of multiple polymorphic genes with risk alleles common in the population, each one determining a moderate portion of the risk [3]. This non-Mendelian pattern of transmission is not exclusive to MS but is shared with other autoimmune diseases and chronic disorders such as type II diabetes and obesity. These conditions are collectively known as complex genetic disorders, which are characterized primarily by polygenic risk and intricate GxE interactions. In the early 2000s, the introduction of chip-based technologies with the capacity to genotype simultaneously hundreds of thousands of SNPs (Single Nucleotide Polymorphisms) allowed the development of a new analytical methodology known as genome-wide association study (GWAS), a hypothesis-free method in which SNPs spaced across the entire genome are screened for association with a particular trait in case–control datasets composed of genetically unrelated individuals [24]. Since 2007, the increasing size of genetic studies has shown that MS risk is influenced by hundreds of genetic variants, many of which are common across the population, with each variant explaining a small proportion of risk [53].

In 2019, the International Multiple Sclerosis Genetics Consortium (IMSGC) reported the results from latest GWAS [54], comparing the allele frequencies of several million common SNPs across the genome between 47 351 people with MS and a population control group of 68 248 people, mainly from European populations. This study resulted in the identification of 233 distinct risk variants, including 200 autosomal SNPs, one SNP on the X chromosome, and up to 32 statistically independent variants across the broader major histocompatibility complex (MHC) area on chromosome six (**Figure 5, Figure 6**). Among these, one rare and four low-frequency protein-coding alleles were associated with small MS risk.

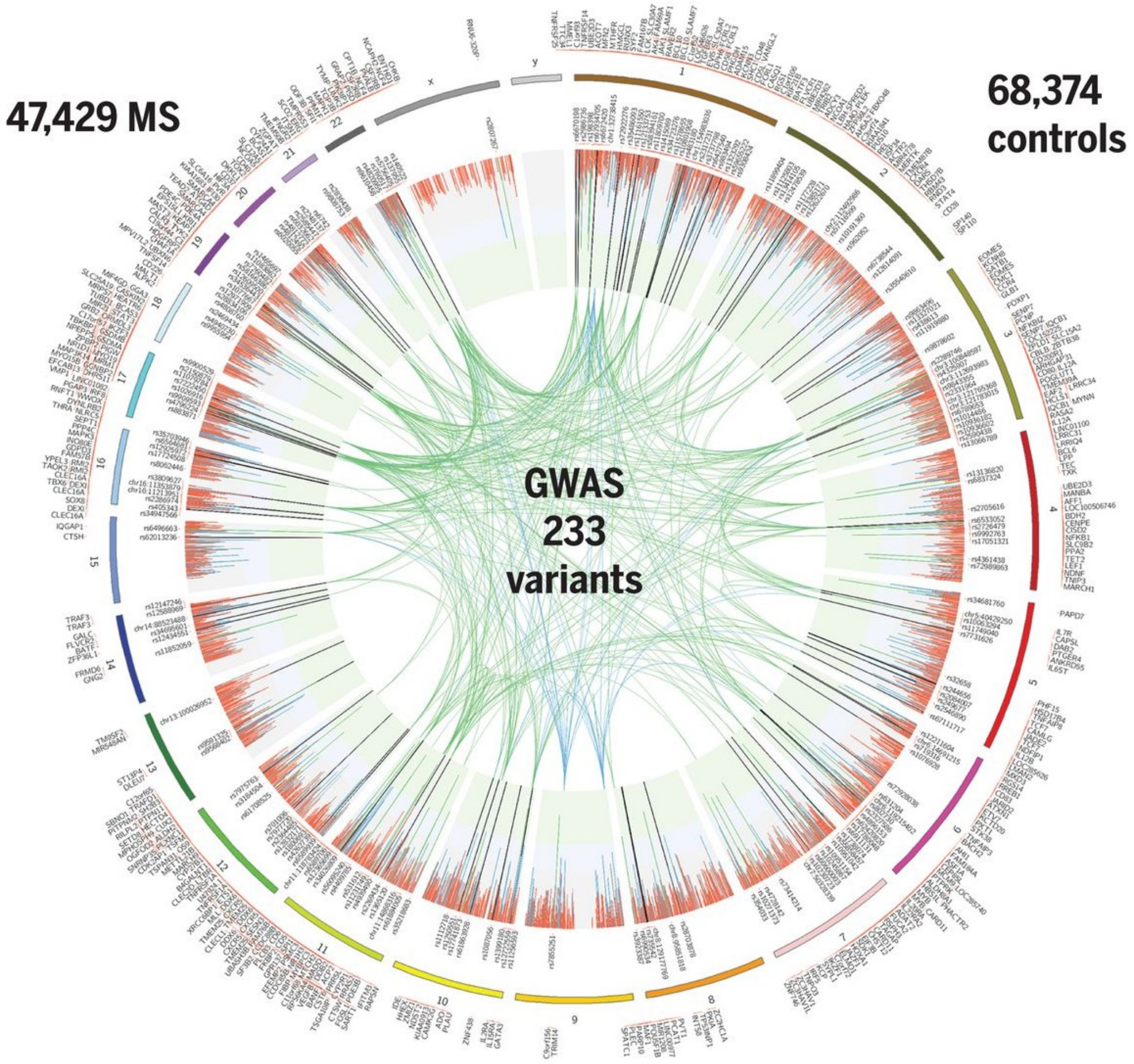


Figure 5. Genomic map of multiple sclerosis susceptibility based on the 2019 International Multiple Sclerosis Genetics Consortium GWAS, highlighting 233 MS-risk variants in the European population. The circus plot summarizes all the Multiple Sclerosis-associated risk loci along with their locus.

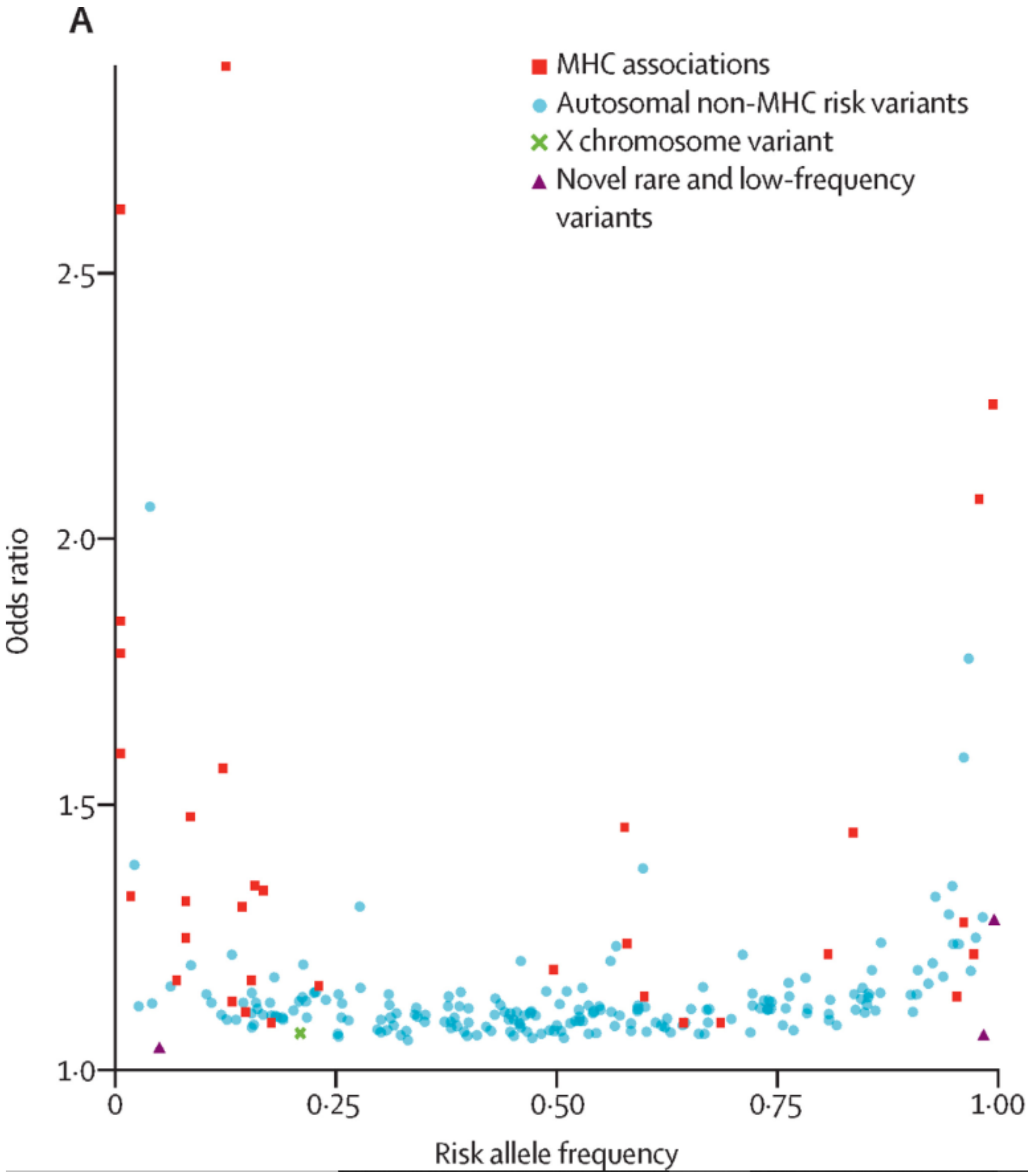


Figure 6. Genomic map of multiple sclerosis susceptibility based on the 2019 International Multiple Sclerosis Genetics Consortium GWAS, highlighting 233 MS-risk variants in the European population based on minor allele frequency and Odds Ratio.

The genetic contribution to the susceptibility of developing MS is then undeniable. However, despite years of research, the biological mechanisms underlying these associations, especially within the MHC region [53], are not fully understood, as no single variant is necessary or sufficient to cause MS; instead, each increases total risk in an additive manner. Moreover, other variants are likely to be associated with MS in different populations and yet to be discovered[3]. According to IMSGC study, the proportion of MS variability explained by their identified genetic variants (over 8 million) was 19.2%, implying that around over 4/5 of MS variability was explained by non-identified genetic variants, environmental factors, and potential GxE and GxG interactions.

Regarding the potential functional mechanism of the identified risk SNPs, common risk variants outside the MHC tend to be in gene promoters and enhancers active in various immune cell types, including natural killer cells, macrophages, microglia, T-cell, and B-cell subsets. These variants are often shared with other autoimmune and inflammatory diseases, suggesting that the disturbances are not specific to the CNS and that MS is mainly an immune mediated disease. This localization of variants implies that MS risk primarily results from subtle changes in gene regulation, leading to alterations in immune cell function that accumulate over time, potentially reaching a pathological state. Among these, approximately 20% of MS susceptibility variants identified by GWASs fall either within or proximal to NFκB signaling genes, including variants proximal to NFKB1 and within TNFRSF1A [55,56]. Previous studies have linked NF-kB activation and JAK-STAT signaling pathways to MS onset [57–60]. In another study [61], a gene network candidate approach has highlighted the putative role of CAMs in MS pathology, possibly connecting the risk to the regulation of BBB crossing by T cells. The challenge in MS genetics is then to translate these discoveries into a deeper understanding of the disease's underlying biology. Moreover, whether these genetic effects interact with environmental factors remains an area of ongoing research, with some indications that specific environmental circumstances might enhance or activate these genotypic effects. To this aim, understanding the mechanisms of MS risk requires the development of new disease models that can quantify the impact of both genetic and environmental effects, as well as their interactions.

Epigenome

It is important to highlight the particular importance of epigenetic modifications, as they could potentially act as a mechanism by which genetic and environmental factors interact. Epigenetic modifications, i.e., heritable changes in gene expression without altering the DNA sequence, can occur through processes like DNA methylation, histone modifications, and microRNA regulation [62]. These modifications, collectively known as the epigenome, are susceptible to influence by environmental factors and are acquired throughout an individual's life, varying among different cell types and tissues. Research has shown that these epigenetic modifications can influence various processes involved in MS pathophysiology, including the breakdown of the BBB, inflammatory responses, demyelination, failure of remyelination, and neurodegeneration [63]. Additionally, investigations into the epigenome are rapidly expanding, revealing connections between epigenetic mechanisms and environmental risk factors for MS. Many of the primary environmental risk factors for MS, such as vitamin D levels, BMI, EBV infection, gut microbiota, diet, and smoking, have been associated with epigenetic modifications [64]. This suggests a potential role for epigenetic changes in the effects exerted by environmental risk factors on the pathogenesis of MS, also explaining gene x environment (GxE) interactions. Although these findings speculate about the involvement of epigenetic modifications in MS, the precise mechanisms and their significance are ongoing areas of research. Understanding these epigenetic alterations in the context of MS could provide valuable insights into disease mechanisms and potential targets for future therapies [65]. Continued research, data collection, and comprehensive epigenome-wide studies are essential to further elucidate these aspects.

In **Figure 7**, a summary for MS risk factors is graphically depicted.

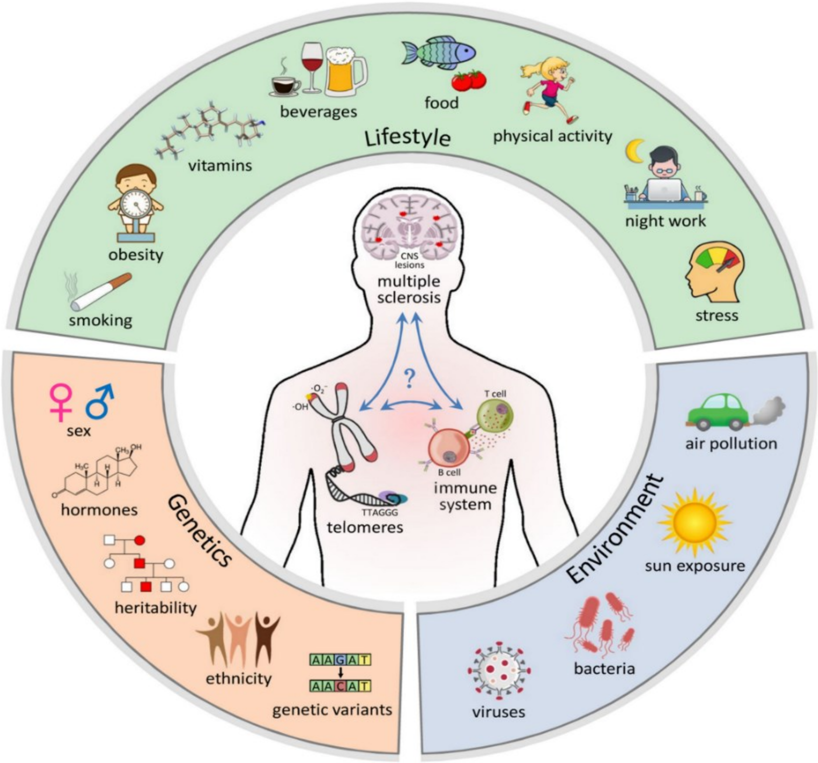


Figure 7. A comprehensive map of Multiple Sclerosis risk factors.

Considering these points, the notion of heritability studies will now be introduced. These studies can assist in quantifying the proportionate influence of genetic and environmental factors in accounting for the variability in MS susceptibility.

1.2 Heritability studies

The concept of heritability refers to the proportion of variations of a phenotypic trait that can be explained by genetic factors [66]. More specifically, according to the additive model [67] the phenotype can be considered as the sum of genetic and environmental effects:

$$\text{Phenotype (P)} = \text{Genetics (G)} + \text{Environment (E)}$$

Where the genetic effects genetic effects (G) variance, con be decomposed in i) additive effects (A), i.e., sum of combined effects of genetic alleles at two or more gene loci, ii) dominant effects (D), i.e., non-additive effects due to the interaction between alleles at the same gene locus, and iii) epistatic effects (V), i.e., non-additive effects due to the interaction between alleles at the different gene loci. The variance of the phenotype (σ_P^2) can be expressed as a sum of unobserved underlying variances:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 = \sigma_A^2 + \sigma_D^2 + \sigma_V^2 + \sigma_E^2$$

Heritability in its broad-sense (H^2) is then expressed by the ratio of the genetic effects variance on the phenotypic variance, i.e., $H^2 = \frac{\sigma_G^2}{\sigma_P^2}$. Instead, the fraction of phenotypic variance owed to genetic additive effects variance alone (σ_A^2) represents the so-called narrow-sense heritability (h^2) [68], which is always less than or equal to H^2. The formula for narrow-sense heritability is then $h^2 = \frac{\sigma_A^2}{\sigma_P^2}$, and, therefore, does not include dominant or epistatic effects. Both measures provide insights into the genetic and environmental architecture of human complex traits and potential ability to dissect out loci associated with trait variation. Thus, determining a high value of heritability is a powerful argument in favour of further research for genetic causes, but it also opens the possibility of predicting heritable risk of illness based on the genetic background. Finally, the ratio $\frac{\sigma_E^2}{\sigma_P^2}$ represents the fraction of phenotypic variance owed to environmental effects (e^2).

In the scientific community there are controversies and misconceptions on heritability interpretation which have been widely discussed, and different authors provided clarifications on its meaning and explanations for its usefulness [66,69–71]. Heritability is a ratio of variances and consequently represents a statistical measure, but it is often misinterpreted causally as the level of causal influence of the genotype on the phenotype. As illustrated by Pearson [66], it's important to view heritability studies as a valuable tool for identifying potential causal factors among the genetic and environmental elements that characterize a population. An accurate interpretation of heritability studies would then be the following:

- $h^2 > e^2$: modifying individuals' genotypes will likely have a greater effect on changing the expression of a trait at the population level compared to controlling environmental factors.
- $h^2 < e^2$: modifying individuals' environmental factors will likely have a greater effect on changing the expression of a trait at the population level compared to controlling the genotype.
- $h^2 \approx e^2$: modifying either individuals' genotypes or environmental factors will likely have a similar effect on changing the expression of a trait at the population level.

Therefore, these studies inform us about the primary factor responsible for trait variation within the population, which inherently sheds light on the underlying causes of that trait. At the individual level, while a heritability estimate greater than zero indicates a causal connection between an individual's genotype and a particular trait, it doesn't provide any direct information about the magnitude of its effect. However, in case $h^2 > e^2$ it is still correct to suggest that an individual's genetic variability contributes more significantly than its environmental variability to cause the deviation of the trait from the population average (and vice versa) [72]. These considerations assume that genetics and environment are statistically independent, and the total variance is solely due to distinct genetic and environmental factors. If this assumption is not met due to covariance between genetics and environment or alterations in causal influence because of GxE interactions, heritability estimates become confounded. It's worth noting that heritability is a measure specific to the local population level and cannot be used to definitively infer the causal role of genetics in determining a trait since populations exhibit variations in genotypes and environmental factors [71]. Nonetheless, the results of heritability analysis can be interpreted within the context of the genetic and environmental background of the specific population under investigation. This contextualization helps identify which factor, whether genetic or environmental variability, plays a more significant role in explaining the expression of the trait in that particular population [69,70,72].

1.2.1 MS heritability in different populations

Quantifying the heritability of complex diseases like MS (OMIM 126200), which is influenced by both genetic and environmental factors, presents a significant challenge due to its unclear etiology [73,74]. Heritability studies serve as a crucial tool for identifying the predominant source of MS variation within a specific population among the potential genetic and environmental causal factors [66,69–71,75]. Traditionally, MS heritability estimates have predominantly relied on the design involving monozygotic (MZ) and dizygotic (DZ) twin pairs [76], as depicted in **Figure 8**.

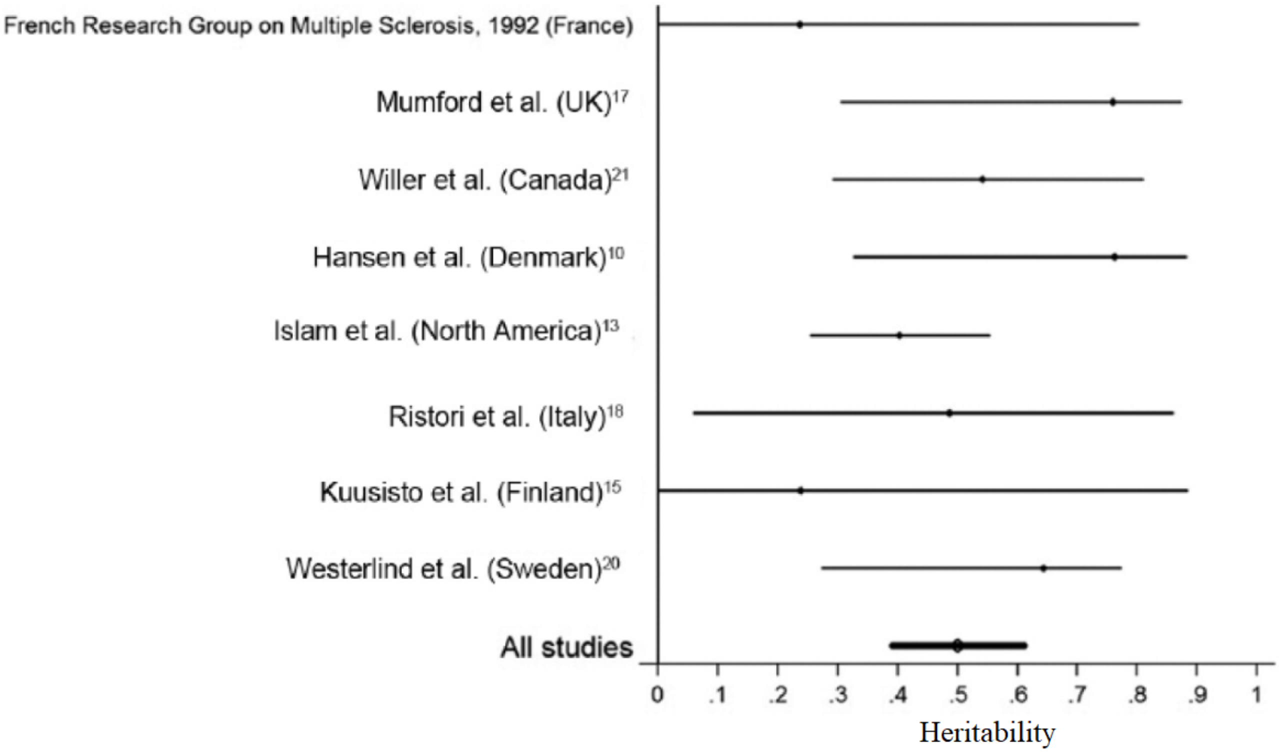


Figure 8. Multiple Sclerosis heritability estimates obtained in different populations using twin design. Source: Fagnani C et al. Twin studies in multiple sclerosis: A meta-estimation of heritability and environmentality. Multiple Sclerosis Journal. 2015;21(11):1404-1413. doi:10.1177/1352458514564492.

These estimates have exhibited some degree of heterogeneity across populations, which was expected, given that each population possesses its unique environmental factors. However, consistently, heritability estimates have remained greater than 0. In countries like the United Kingdom (UK), Denmark, and Sweden, these estimates have been close to 80%, signifying that approximately 80% of the variability in MS within the population is attributed to genetic factors. In these countries, environmental factors play a relatively smaller role in explaining variations in MS expression. Conversely, in France and Finland, the heritability estimate for MS has been around 20%, indicating that certain environmental factors have a more substantial impact on explaining MS variability in these populations. In a hypothetical scenario, this suggests that to reduce MS expression in the UK, Denmark, and Sweden, the primary focus should be on investigating the genetic diversity within the population. In contrast, to reduce MS expression in France and Finland, efforts should concentrate on identifying and mitigating the environmental factors contributing to the disease. Understanding whether genetic or environmental factors predominantly drive MS expression in a given population can provide valuable insights into identifying the key causal factors, thereby enhancing disease prevention strategies.

1.3 Aim of the research

1.3.1 MS heritability in the Sardinian population

Up to this point, there have been no efforts to estimate heritability in the Sardinian population due to the relatively low prevalence of multiple sclerosis (MS) and the limited population size of the Sardinian Island (1,611,621 inhabitants as of the 2019 census [77,78]). This limitation makes it challenging to gather a sufficiently large number of twin pairs to derive an informative and accurate heritability estimate [76,79]. One way to partially address this challenge is to include individuals with various familial relationships, which increases the sample size and consequently enhances statistical power. However, this approach becomes less effective as more distant relatives are included, primarily because they share a smaller proportion of their genetics [80]. Importantly, this design doesn't necessarily require genotype data, as the genetic relatedness between individuals can be estimated based on expected relationships [68]. Additionally, pedigree-based studies offer advantages compared to twin studies, as they generate heritability estimates that are less influenced by potential shared environmental effects, i.e., environmental influences that make individuals raised in the same environment more similar to each other [81,82]. Consequently, heritability estimates derived from pedigree-based studies tend to be more robust when there are model misspecifications regarding shared environmental effects [83].

In this study, the main objective was to explore the variability in MS expression within the Sardinian population by quantifying the relative contributions of genetic variability (h^2) and environmental factors. A sample made of 24 Sardinian extended families, identified through MS affected probands in the Nuoro province, was then considered to investigate this aim. Additionally, due to the historical depth of the available family data, it was possible to explore the role of environmental factors over time, including shared environmental effects, individual environmental effects, sex, and year of birth. The consideration of year of birth is significant as it can serve as a proxy for various early environmental exposures which changed over time, especially after the post-World War II industrialization, i.e., socioeconomic factors, dietary habits, lifestyle, and sanitary conditions (referred to as the "Westernization process")[84–87]. Furthermore, it accounts for the potential impact of the malaria eradication program conducted from 1946 to 1950, which involved the use of the insecticide DDT (dichloro-diphenyl-trichloroethane) [88]. **Figure 9** illustrates an estimate of malaria occurrence in Italy in 1932, as provided by the Istituto Superiore di Sanità. It specifically emphasizes the Sardinia region as having a notable presence of malaria.

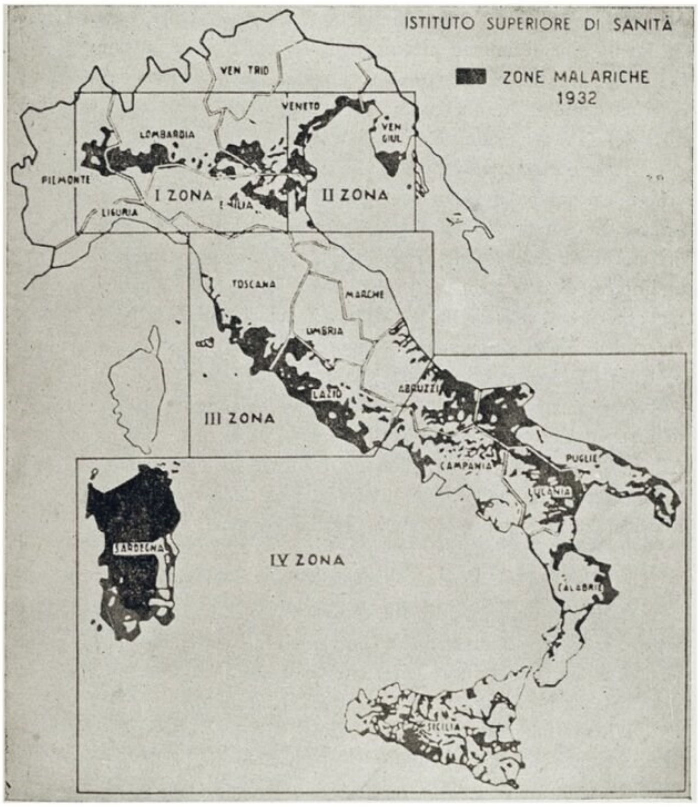


Figure 9. Malaria distribution in Italy as recorded in 1932 by the Istituto Superiore di Sanità.

These aspects could be linked to the constant MS incidence observed since the 1950s in the Nuoro province [84] and other Sardinian provinces [89]. Different authors have also underlined how a better diagnostic accuracy cannot fully account for this steady increase in MS [84,85,89,90], since the magnitude of this trend has not been observed in any other Italian areas during the same period. As a comparison, **Figure 10** shows the MS incidence rates in the Sardinian region of Nuoro and the continental Italian province of Ferrara, located in the Emilia-Romagna region, between 1965 and 1995.

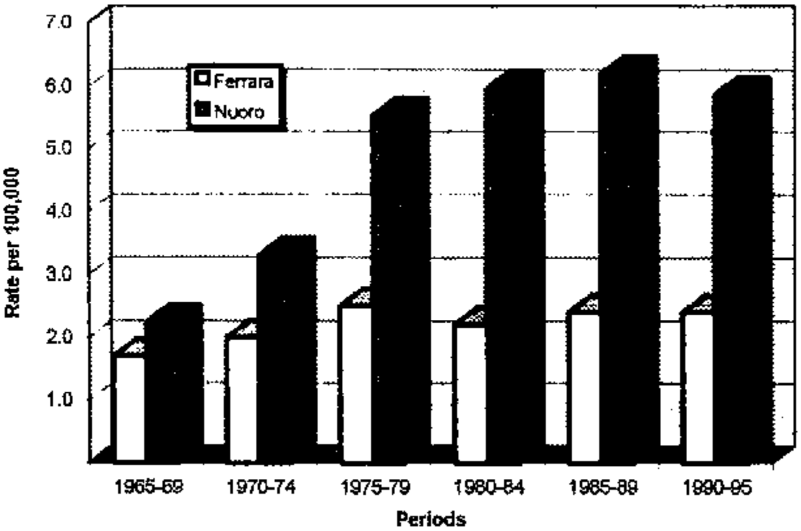


Figure 10. Multiple Sclerosis incidence over time in the Italian provinces of Nuoro (black) and Ferrara (white).

Notably, it demonstrates that MS incidence in the Nuoro province doubled that of the Ferrara province within only a decade. This raises inquiries about whether sudden adverse environmental changes are responsible for the increase in MS incidence. The previously mentioned malaria eradication has been linked in literature to this dramatic increase in MS incidence under the light of the hygiene hypothesis. This hypothesis suggests that reduced exposure to infectious agents, due to environmental or lifestyle changes, is associated with a higher prevalence of allergic hypersensitivity and autoimmune disorders [91]. In this context, the reduced exposure to the infectious agent is represented by the eradication of *Plasmodium falciparum* [92]. Studies have linked the presence of the A30-B18-DR3 HLA haplotype (strongly associated with MS in Sardinia) to high malaria prevalence areas [93]. However, this hypothesis involves complex interactions between the host's immune responses, characteristics of invading microorganisms, environmental exposures, and genetic factors. The availability of the 24 extended Sardinian families, along with the information on the year of birth, allows to obtain more insights concerning this hypothesis.

Considering the significant impact of MS on the affected individuals' well-being, as well as its broader implications for public health and the economy, there is an urgent and compelling need to uncover the fundamental cause of the disease. This involves identifying the factors that have a causal connection to MS. Such insights can subsequently inform appropriate treatment strategies, guide preventive measures, and lay the foundation for future advancements in precision medicine. In summary, given the intricate nature of the disease and the environmental changes within the Sardinian population, exploring MS susceptibility variability within these extended families holds the potential to make a meaningful contribution to the research on the causal determinants of MS in this specific population [66,75].

1.3.2 The limitations of available methodologies to estimate heritability

Several methodologies have been proven to give unbiased heritability estimates for binary traits when using families that have been randomly selected from the population [94]. When dealing with low-prevalence diseases, like MS, family members are included in the study as relatives of an already enrolled affected member (proband). This a common type of sampling in genetic studies, but unfortunately leads heritability estimates to be affected by ascertainment bias. This bias artificially inflates the additive genetic effects due to an overrepresentation of affected cases compared to the general population [77,95,96]. To address this issue, Kim, Kwak and Won [96] introduced a liability threshold model for binary traits (LTMH), suitable for families ascertained from a proband. LTMH helps estimate heritability on a liability scale while correcting for ascertainment bias. The authors assessed the performance of this method, which relies on the Expectation-Maximization (EM) algorithm, through simulations involving 500 nuclear families sampled from affected probands and varying disease prevalence (e.g., 5%, 10%, 20%). In all scenarios, LTMH provided accurate heritability estimates on the liability scale, effectively correcting for ascertainment bias. However, using the EM algorithm for heritability estimation has certain drawbacks, such as lacking a precision measure for statistical inference, i.e., standard error and related confidence interval, and computational inefficiency when dealing with extended families or more complex statistical models able to correct for shared environment effects. These limitations significantly restricted the capacity of the LTMH method to deliver comprehensive results. Additionally, due to its computational inefficiency when applied to large extended families, it was unsuitable for use with the Sardinian extended family sample.

1.3.3 Main aspects of this research

Given the considerations provided in the previous paragraphs, the main objectives of this thesis project could be summarized as follows:

1. To address the limitations of the EM algorithm in LTMH affecting computational efficiency, model flexibility and statistical inference. This involved developing a new statistical framework, to be empirically evaluated for its accuracy, precision, and computational efficiency. Furthermore, the developed framework could be applied in future research to analyse other low-prevalence complex traits, aiming to quantify the contributions of genetic and environmental factors in specific populations of interest.
2. To apply the newly developed statistical framework to the dataset of Sardinian extended families. My goal was to identify the primary factors contributing to the variability in MS susceptibility within this population. These factors comprised genetic variability, environmental factors (both individual and shared among relatives), sex, and year of birth. The findings may provide insights into which specific factor requires in-depth investigation to elucidate the high MS incidence observed in the Sardinian population.

2 METHODS

2.1 Bayesian-LTMH

2.1.1 Addressing the limitations of EM algorithm for heritability estimation

As mentioned in the Introduction, to estimate the heritability of MS in the Sardinian population, using a sample of 24 Sardinian extended families ascertained from affected probands [97], the LTMH method developed by Kim, Kwak, and Won [96] was initially considered. Making use of family-based samples ascertained from a proband, the method utilizes a liability threshold model designed to partition the variability of a binary trait into genetic and environmental effects, adjusting estimates from ascertainment bias. To estimate the model's parameters, the method employs the EM algorithm [98]. EM algorithm is an iterative method based on a frequentist approach (opposed to Bayesian statistics), as it is primarily used to find (local) maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. Briefly, in the Expectation step (E-step), the algorithm starts with an initial guess for the unobserved model's parameters. It computes the expected values of the unobserved or missing data given the observed data and the current parameter estimates. In the Maximization step, the algorithm updates the model's parameters to maximize the likelihood of the observed data. It uses the expected values computed in the E-step as if they were actual data points with known values. The parameters are adjusted to improve the fit of the model to the observed data. These two steps are repeated iteratively until the algorithm converges to a set of parameter estimates that maximize the likelihood of the observed data. In each iteration, the parameter estimates are refined, and the likelihood of the data is improved. In summary, the EM algorithm is particularly useful when dealing with complex models with hidden or unobservable variables. However, the usage of EM algorithm for parameter estimations in LTMH leads to the following limitations:

a) The EM algorithm does not directly provide a precision measure, i.e., standard error, for parameter estimates, as it would necessitate complex analytical calculations [99,100]. This issue leads to difficulty to conduct statistical inference.
b) Even if standard errors are calculated, determining how to compute confidence intervals for heritability is challenging. Assuming a normal distribution for heritability can be misleading as confidence limits could go below 0 and/or above 1, while the parameter is characterized within the [0,1] bounds [101].
c) The conditional E-step of the EM algorithm involves computationally intensive tasks, such as estimating first and second moments of multivariate truncated normal distributions. This leads to slow computational efficiency and issues in obtaining algorithm's convergence, especially when i) dealing with extended families beyond nuclear families and ii) incorporating additional variance components, e.g., to consider shared environment effects.

These limitations hold particular significance in the context of family-based and complex disease studies, prompting me to consider the EM algorithm as an inefficient method for estimating MS heritability in my research. Moreover, when implementing LTMH using the EM algorithm on the Sardinian sample, the algorithm encountered difficulties and remained stuck after the first iteration. Therefore, my objectives were as follows:

i. To provide a precision measure for the parameters estimates, facilitating statistical inference and addressing limitation (a).
ii. To establish an interval range for heritability estimates that naturally reside within the [0,1] bounds, addressing limitation (b).
iii. To retain the extended structure of the Sardinian families enhancing the computational efficiency of the LTMH method and overcome limitations (c-i).
iv. To incorporate adjustments for heritability estimates to account for shared environmental effects, thereby implementing a more complex model and addressing limitations (c-ii).

To overcome these limitations, Bayesian statistics and Markov-Chain Monte Carlo (MCMC) techniques were explored as alternative methods for heritability estimation. The rationale behind this choice is as follows:

- Bayesian statistics offers posterior distributions for the parameters of interest, effectively addressing limitations (a) and (b).
- MCMC methods have demonstrated their speed and efficiency, particularly when dealing with statistical models containing numerous unobserved variables that result in likelihood functions featuring multiple integrals [102]. Using these methods, there's no need to explicitly compute the integrals in the likelihood function, and the unobserved variables can be sampled alongside the model parameters. This would help to address limitation (c).

In the following paragraphs, Bayesian statistics will be discussed, along with its principles of inference and the potential advantages it offers for parameter estimation compared to the EM algorithm-based approach.

2.1.2 Bayesian statistics and MCMC methods

2.1.2.1 Bayes' theorem

In contrast to frequentist analyses, Bayesian statistics relies on the application of Bayes theorem [103]:

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)} * \mathrm{P(A)}}{\mathrm{P(B)}}$$

Where A and B are events, and $\mathrm{P(A)}$ and $\mathrm{P(B)}$ their respective marginal probabilities. $\mathrm{P(B|A)}$ represents the probability of observing the event B conditional to observing event A. This latter also corresponds to likelihood of A given a fixed event B, i.e., $\mathrm{L(A|B)}$. Finally, $\mathrm{P(A|B)}$ represents the probability of observing the event A conditional to observing event B, and it is called the posterior probability of A given B. $\mathrm{P(A|B)}$ can be interpreted as the degree of belief in A after incorporating information from B (see **Figure 11**).

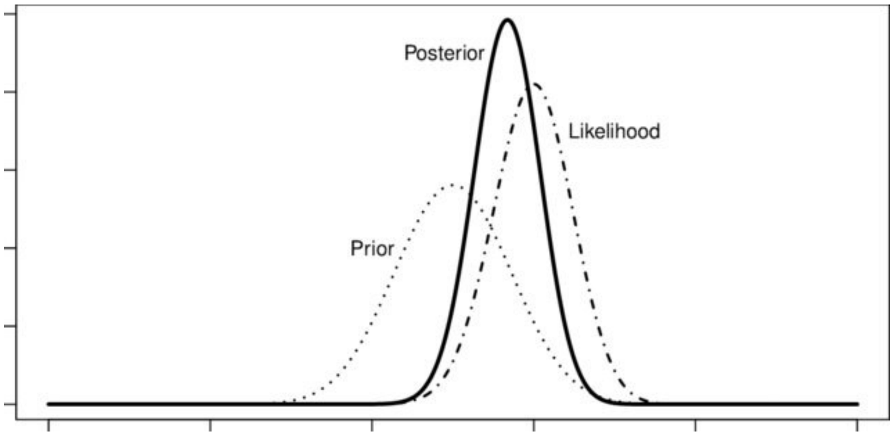


Figure 11. Example of posterior distribution resulting from Bayes' theorem incorporating prior information in the likelihood function.

Bayesian methods offer a framework for incorporating prior information, denoted as $\mathrm{P(A)}$, into the analysis. This prior information represents the initial level of belief in the event A. As new data is observed, the prior belief gets updated, contributing to the quantification of evidence supporting a hypothesis. Although some criticize this approach as a potential source of bias or excessive subjectivity into results, judiciously incorporating prior information can regularize computations and enhance the stability of statistical inferences. There exist various approaches to construct a prior distribution, which can stem from historical data like previous experiments or be elicited through the subjective judgment of an experienced expert. In cases where no information is available, one may opt for weakly informative or non-informative priors, which primarily serve to constrain inferences within a reasonable range without introducing specific knowledge about A from prior sources.

Another important aspect of Bayesian statistics, compared to the frequentist approach, is that inferences are made directly about the parameters of interest. Parameters, and functions of them, have probability distributions, so it is possible to make statements such as "the probability that the parameter is greater than 1 is 0.80." [104].

2.1.2.2 Bayesian modeling outline

The process of fitting a Bayesian model generally involves the following steps [105]:

1) Initially, a model is defined based on the hypothesis the researcher wants to test. This model includes a parameter (or in more complex cases, multiple parameters) to estimate, along with a specification for the data distribution (e.g., normal distribution).
2) Subsequently, a prior distribution is established to represent previous beliefs regarding plausible values of the parameter. In contrast to frequentist statistics, which treat the parameter as fixed, Bayesian methodology treats it as random, reflecting the previous beliefs prior to data examination. The prior distribution is characterized by the so-called hyperparameters, which are defined by the research to obtain the desired distribution. Alternatively, it is also possible to treat hyperparameters as random variables defined by hyperpriors distribution. The use of a hyperprior then allows one to express uncertainty in a hyperparameter.
3) The next step involves combining these prior beliefs with information derived from the observed data, resulting in the formation of a posterior distribution. This posterior distribution

characterizes the updated beliefs about the probable values of the parameter after observing the data.

4) Finally, the posterior distribution is summarized in a suitable manner, typically by calculating position measures, e.g., mean, median, mode, and dispersion measures, e.g., standard deviation (SD). Moreover, a credible interval is often calculated so that the parameter is included in a particular range with a specified probability, e.g., a 90% credibility interval conveys the information that the true parameter is contained within that interval with a probability equal to 0.9. Additionally, other informative metrics can be computed from the posterior distribution, such as the probability that the parameter is greater/lower than a certain value of interest. The posterior distribution then allows to obtain more information compared to the frequentist estimates and related confidence intervals.

2.1.2.3 Posterior distribution

Following Bayes' theorem, exact posterior distribution can be obtained under certain conditions and in specific situations. This typically occurs in relatively straightforward and well-studied statistical models. Denoting with $x = (x_1, \dots, x_k)$ a vector of observations with underlying statistical distribution function of parameters $\theta = (\theta_1, \dots, \theta_k)$, the posterior distribution's density is proportional to the product of the likelihood function for the observations, i.e., $l(\theta|x)$, and the prior distribution, i.e., $P(\theta)$:

$$P(\theta|x) \propto l(\theta|x)P(\theta)$$

In cases where the prior distribution and likelihood function belong to the same parametric family and have a specific mathematical relationship, known as conjugacy, the posterior distribution is also in the same parametric family as the prior. Common examples include the normal distribution with known variance, where the posterior for the mean is also normal, or the Beta-Binomial model, where the posterior for the success probability is also a Beta distribution. Conjugate priors simplify the calculation of the posterior distribution.

2.1.2.4 MCMC methods

In many cases, computing the exact posterior distribution, which summarizes what we know about the parameters after considering both data and prior beliefs, is analytically intractable. This is the case when the moments of the posterior distribution require several integral calculations which often cannot be solved analytically. In these cases, a solution can be obtained approximating the posterior distribution based on asymptotic results. MCMC methods provide a way to approximate the posterior distribution by generating a sequence of its samples [102]. These samples are produced through a Markov chain, which is a sequence of random variables where each variable depends only on the previous one. Practically, an ensemble of chains is generally developed, starting from a set of points arbitrarily chosen and sufficiently distant from each other. These chains are stochastic processes of "walkers" which move around randomly according to an algorithm that looks for places with a reasonably high contribution to the integral to move into next, assigning them higher probabilities. These samples can be used to evaluate an integral over that variable, as its expected value or variance [106].

2.1.2.5 Metropolis-Hastings algorithm and Gibbs sampling

To implement MCMC methods, several algorithms have been developed. Metropolis-Hastings algorithm, represents the first widely used MCMC method, which uses a proposal distribution to generate new samples, and involves the following steps [107]:

1) Initialization: start of a Markov Chain with an initial value for the parameters of interest, i.e., θ^0.
2) Proposal: a new set of proposed parameter values θ^* is sampled using a proposal distribution q, i.e., $q(\theta^*|\theta^{(t-1)})$.
3) Acceptance: decide whether to accept the proposed values based on a probability ratio:

$$R(\theta^*|\ \theta^{(t-1)})\colon \frac{P(\theta^*|x)q(\theta^{(t-1)}|\theta^*)}{P(\theta^{t-1}|x)q(\theta^*|\theta^{(t-1)})}$$

By applying Bayes' theorem for the posterior probability terms in the formula above we get:

$$R(\theta^*|\ \theta^{(t-1)})\colon \frac{P(\theta^*)l(\theta^*|x)q(\theta^{(t-1)}|\theta^*)}{P(\theta^{t-1})l(\theta^{t-1}|x)q(\theta^*|\theta^{(t-1)})}$$

Where $R(\theta^*|\ \theta^{(t-1)})$ is defined as Metropolis-Hastings ratio. This ratio provides a probability defined as:

$$\alpha = \min\{1, R(\theta^*|\ \theta^{(t-1)})\}$$

To establish the acceptance of θ^*, a uniform random number $u\ \in [0,1]$ is generated.
If $\alpha > u$ then the proposed values θ^* are accepted, i.e., $\theta^t = \theta^*$. Conversely, θ^* is rejected, and $\theta^t = \ \theta^{(t-1)}$. Depending on the result, the new values θ^t become the current state of the chain.
4) Repeat: continue the process for a predefined number of iterations.
5) Results: over time, this process generates a sample from the target posterior distribution. It updates one parameter at a time while keeping others fixed, making it particularly effective when the posterior distribution can be expressed conditionally.
6) Burn-in and Thinning: MCMC chains often start with an initial "burn-in" phase to allow the chain to reach the target distribution and remove any initial bias. Therefore, a pre-specified initial number of sampling iterations is discarded. Moreover, to reduce autocorrelation between close iterations, a subsampling of the generated samples is retained; this practice is defined as thinning.
7) Convergence and Mixing: assessing convergence is crucial to ensure that the MCMC chain has explored the entire posterior distribution. Diagnostic tools like the Gelman-Rubin statistic and visual inspection of trace plots are used to evaluate convergence.

A graphical exemplified depiction of the algorithm is exemplified in **Figure 12**.

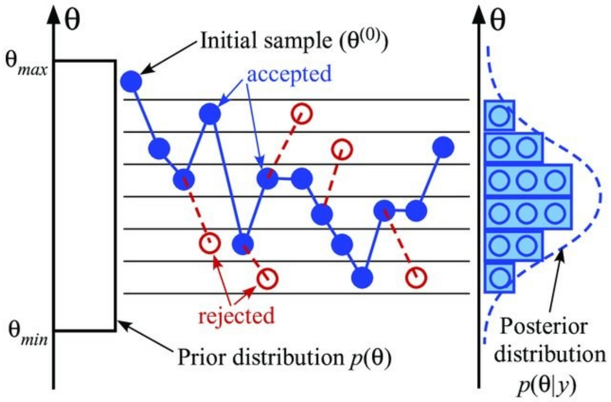


Figure 12. Graphical depiction of Metropolis-Hastings algorithm in obtaining samples from the posterior distribution.

Gibbs sampling is another MCMC method which is a special case of Metropolis-Hastings. It is particularly suitable when dealing with multivariate distributions, as it iteratively updates one parameter at a time while keeping others fixed sampling from the conditional distribution rather than to marginalize by integrating over a joint distribution [108].

2.1.2.6 Hamiltonian Monte Carlo algorithm

Among the latest developed MCMC algorithms, Hamiltonian Monte Carlo (HMC) represents an advanced method for efficient sampling of complex posterior distributions, especially in high-dimensional spaces [109]. It is inspired by the principles of classical mechanics, particularly Hamiltonian dynamics, where particles move through a physical space subject to potential and kinetic energy functions [110]. The system evolves over time, simulating the dynamics, or trajectories, of particles in a potential energy landscape. HMC exploits these concepts to enhance the exploration of the parameters space, i.e., to sample parameters from the posterior distribution in the Bayesian inference context [111]. HMC combines random walk exploration with these deterministic trajectories to propose new states efficiently. Each state in the Markov Chain is represented by a pair of values: the position and their associated auxiliary momenta. The position represents the parameters of the target posterior distribution that we are trying to sample from, in other words the model parameters in a Bayesian inference problem. The positions represent the current values of these parameters in the parameter space. On the other hand, the momenta (plural of momentum) are auxiliary variables introduced in HMC and associated with each parameter. The momenta are not parameters of the model but rather help simulate the dynamics of the system. Each momentum is associated to one parameter and is used to create a joint state with the parameter value. Therefore, the combination of parameters and momenta constitutes the state of the trajectory, defining a point in the joint space of parameters and momenta. The goal of HMC is then to simulate trajectories in this joint space over time, guided by the Hamiltonian dynamics equations, which describe how the positions, i.e., parameters, and momenta change. To make these equations computationally tractable, they are discretized using a numerical integration scheme, such as the leapfrog method [112]. This method represents an approximate solution to the motion of non-interacting classical particles. A step size, denoted as ε, determines the size of the discrete time steps taken during the numerical integration of the equations. Smaller ε values result in more accurate but computationally intensive simulations, while larger ε values speed up the computation but may lead to less accurate trajectories. The numerical integration of these equations generates a proposed state (position and momenta) in the joint space, which is then subject to acceptance or rejection based on the Metropolis-Hastings

criterion based on the ratio of posterior probabilities. If accepted, the positions (parameters) become the new values of the Markov Chain, while the momenta are discarded. Thus, while the momenta are auxiliary variables introduced during the simulation, the final output of HMC is a sample from the posterior distribution, which includes the parameter values.
As for other MCMC algorithms, HMC requires the user to operate a tuning of the algorithm's parameters. Other than the specification of number of burn-in and sampling iterations which is common with other MCMC methods, HMC requires to set the step size ε and the number of leapfrog steps for numerical integration. However, a variant of HMC, the No-U-Turn Sampler (NUTS) [113], has the great advantage to automatically tunes the step size ε and the number of leapfrog steps required for each trajectory. The step size ε is adaptively adjusted during the trajectory simulation to account for the curvature of the target distribution, helping to avoid issues like poor exploration or numerical instability.

2.1.2.7 Advantages of using Bayesian inference based on HMC algorithm.

In summary, the HMC algorithm, along with its variant NUTS sampler, combines principles from classical mechanics with Bayesian inference to efficiently navigate complex posterior distributions. This makes it an invaluable tool for Bayesian analysis, especially in scenarios with high-dimensional parameter spaces. When compared to the Metropolis-Hastings algorithm, the NUTS sampler offers several advantages [111]:

1. Enhanced Efficiency: NUTS excels at exploring high-dimensional parameter spaces more efficiently. It's particularly suitable at traversing these spaces without the inefficiencies associated with random-walk-like behavior, which often necessitates a large number of samples to achieve convergence.
2. Reduced Random Walk: by minimizing the random-walk-like behavior, NUTS helps reduce the number of samples required for the Markov chain to reach convergence. This improvement in exploration efficiency is especially advantageous when dealing with complex models.
3. Improved Handling of Correlations: NUTS handles correlated parameters more effectively. It helps mitigate the challenges posed by parameter dependencies, leading to a more accurate and efficient exploration of the joint posterior distribution.

As a result, HMC methods, grounded in Bayesian inference principles, serve as a potent approach when analytical solutions are impractical, or frequentist methods struggle to converge or grapple with computational limitations [113]. These advantages led me to consider NUTS as a valid alternative to the EM algorithm to estimate parameters in the LTMH method. In the next paragraphs it will be discussed how Bayesian statistics and NUTS fit in the context of LTMH and heritability estimation.

2.1.3 Notations and LTMH specification

LTMH methodology will now be described along with further insights on modeling features necessary for heritability studies. A sample is considered, made of N individuals, denoted with j, clustered in F families denoted with i. The N observed binary phenotypes Y_{ij}, which can be considered as cases and controls, are determined by unobserved continuous liability scores L_{ij} and a fixed threshold c [114]. The threshold c is placed on the liability distribution so that the portion of the distribution equal to the trait prevalence in the population is above the threshold [115]. LTMH can include covariates, i.e., fixed effects, which are useful used to adjust for potential confounders effects and to improve model fitting leading to higher estimates precision [116]. Moreover, it may also be of interest measuring their explained proportion of phenotypic variance on a liability scale. Covariates, which are assumed to be standardized so that they are centered to their mean value, are denoted with a matrix X_{NxB}, where B denotes the number of covariates, while b represents the respective vector of fixed effects parameters $\beta_1, \dots, \beta_B$. The liability scores L are assumed to be distributed following a multivariate normal distribution, i.e., $L \sim MVN(Xb, \Sigma)$. The intercept term is constrained to 0, which means that $L = 0$ when covariates are equal to their mean value. The covariance matrix Σ is composed of a block diagonal matrix consisting of the covariance matrices Σ_i calculated within each family.

In the development of the LTMH method under a Bayesian framework, the interpretation of the components included in Σ, aiming to obtain an unbiased heritability estimate, will now be discussed. Following the standard polygenic additive model [67], assuming null epistatic and GxE effects, Σ is defined equal to the sum of genetic and environmental effects variance components:

$$\Sigma = \sigma_A^2 K + \sigma_D^2 V + \sigma_C^2 H + \sigma_E^2 I$$

Where σ_E^2 represents the individual environmental effects (E) variance and I the respective identity matrix, σ_A^2 represents the additive genetic effects (A) variance and K the respective kinship matrix multiplied by two, meaning that components Φ_{jk} are defined as the probability, at a given locus, that two gene copies chosen at random from two individuals j and k are identity-by-descent (IBD). σ_D^2 represents the dominant genetic effects (D) variance and V the respective dominant genetic matrix whose components v_{jk} are defined as the probability at a given locus that two individuals j and k share two gene copies IBD [117]. If σ_D^2 is incorrectly assumed to be null then σ_A^2 could result in an inflated estimate by a term equal to $0.5\sigma_D^2$ [81,118]. Shared environment effects (C) variance σ_C^2 along with the respective H correlation matrix, are also modeled to avoid an inflation of σ_A^2 estimate due to common environmental influences potentially resulting in phenotypic resemblances [82]. It is not an easy task to specify to what extent or in what form these environmental effects exert their presence through familial correlations and from one generation to another [119]. Reasonably, the main source of shared environment effects can be identified in siblings/twins raised in the same household, as these likely share the same eating habits, family lifestyle, infectious diseases and sources of pollution [120–122]. Therefore, H matrix components h_{jk} can be defined as equal to 1 for groups of siblings. It is also important to notice that in this case it results impossible to disentangle dominance genetic and shared environment effects, as the former effect σ_D^2 would be completely masked by the latter σ_C^2 [123]. Still, using an ACE model, i.e., a model which jointly consider additive genetic effects (A), shared environment effects (C) and individual environment effects (E), σ_C^2 estimate would adjust for σ_D^2 avoiding σ_A^2 inflation [81,124], but it is expected to be inflated by a term equal to $0.25\sigma_D^2$. Since the principal aim is to accurately estimate σ_A^2, an ACE model will be considered acknowledging the potential inflation of σ_C^2 due to dominant effects. Nevertheless, it is also worth mentioning that dominant effects have been shown to have little impact on complex traits [125]. Additionally, one can model the variance of environmental effects shared by siblings which are "transmitted" from the parents to the offspring; in fact, it is likely that there is some dependence between parents' and offspring's environments in which they were raised; for instance, if parents grow up in poverty, their children are more likely to grow up in poverty as well [83]. This similarity in environments is difficult to model because parents and offspring do not grow up in the same household in the same way as siblings do, and any assumption on the similarity in environment would be speculative unless explicitly assessed [83]. Therefore, it is reasonable to assume that only a certain fraction of this shared environment is "transmitted" from a parent to the offspring. For a nuclear family, e.g., mother, father and two children, shared environmental effects, including a transmitted effect from the mother to the offspring, can be modeled specifying the following covariance matrix:

$$\sigma_C^2 H = \begin{bmatrix} \sigma_C^2 & 0 & t\sigma_C^2 & t\sigma_C^2 \\ 0 & \sigma_C^2 & 0 & 0 \\ t\sigma_C^2 & 0 & \sigma_C^2 & \sigma_C^2 \\ t\sigma_C^2 & 0 & \sigma_C^2 & \sigma_C^2 \end{bmatrix}$$

With t being an unknown parameter bounded between 0 and 1. Instead of treating t as an unknown parameter to be estimated, Σ can be formulated as follows [119]:

$$\Sigma = \sigma_A^2 K + \sigma_C^2 H_1 + t\sigma_C^2 H_2 + (\sigma_E^2 - t\sigma_C^2) I$$

Where H_1 and H_2 represent, respectively, the correlation matrix with values equal to 1 between siblings and between the mother and the offspring. Therefore, the variance of the effects due to shared environment between siblings σ_C^2 and the effects of the correlated environment shared between siblings transmitted by a parent $t\sigma_C^2$ can be estimated. The sum of σ_C^2 and $t\sigma_C^2$ could then be considered as an estimate of the total shared environment effects variance. Using extended families has the advantage, over trios/nuclear families, to maintain model parameters identifiable in case of several different shared environmental components are added to the model, e.g., transmitted shared environmental effects from each parent can be modeled separately without losing parameters identifiability; this is due to the larger number of degrees of freedom made available by the increased variety in genetic relationships [126,127]. Moreover, the variance of shared environment effects between spouses can also be considered, as previous literature showed how partners of people with specific diseases are at increased risk of the disease themselves, sharing similar lifestyle and behaviors being members of the same household [128]. To check for parameters identifiability one can verify if the S covariance matrices added in Σ are linearly independent, meaning that the equation $x_1 M_1 + \cdots + x_S M_S = 0$, where M_i and x_i represent the correlation matrix and the respective variance as specified in the model, can be solved only for $x_1 = \cdots = x_S = 0$. Therefore, the choice in modeling the shared environmental effects is somewhat arbitrary, as it depends on the pedigree structure other than the knowledge about the trait of interest [122]; however, it represents a fundamental feature to avoid inflated heritability estimates. Now that all the components in Σ have been defined, the sum of the variance components, which is equal to the phenotype liability variance σ_L^2, is assumed to sum to 1 to avoid parameters identifiability problems [129]. The individual environment variance σ_E^2 is then derived as the complementary term with respect to all other variance components in the model, e.g., $\sigma_E^2 = 1 - \sigma_A^2 - \sigma_E^2 - t\sigma_C^2$. It should be noticed that since a fraction of phenotypic variance, on liability scale, is explained by the fixed effects b, the obtained parameters posterior distributions should be interpreted conditional to the proportion of phenotypic variance explained by the B covariates included in the model, i.e., σ^2_β. As described by Villemereuil et al. [116], σ^2_β should be considered part of total phenotypic variance, i.e., $1 + \sigma^2_\beta$, where $\sigma^2_\beta = \text{var}(Xb)$, to obtain a correct and marginal interpretation for the parameters. Therefore, it is necessary to derive the ratio between σ^2_A, σ^2_C, σ^2_E and σ^2_β and $1 + \sigma^2_\beta$ to obtaining the marginal posterior parameters distribution for:

- h^2: narrow-sense heritability, i.e., proportion of phenotypic variance, on liability scale, explained by additive genetic effects.
- c^2, proportion of phenotypic variance, on liability scale, explained by shared environment effects.
- e^2, proportion of phenotypic variance, on liability scale, explained by individual environment effects.
- τ^2_β, proportion of phenotypic variance, on liability scale, explained by covariates.

This is only relevant when covariates are included in the model, otherwise σ^2_A, σ^2_C, and σ^2_E parameters would automatically be equal to h^2, c^2, and e^2 as the total phenotypic variance is equal to 1 for construction.

2.1.4 Modeling GxE interaction effects variance

LTMH can be expanded to investigate the role of GxE interaction effects, which is particularly important when environmental factors are considered. GxE effects imply that genetic variants have varying causal effects on outcomes depending on the environmental conditions [130]. Understanding GxE effects helps clarify why individuals with the same genetic predisposition may experience different outcomes when exposed to different environmental factors. Assessing GxE effects is crucial to test the validity of the additive principle, as their presence could confound h^2 estimates. GxE effects, between genetic variability and an environmental factor, can be studied and modeled using the approach described by Almasy and Blangero [131], defining Σ as follows:

$$\Sigma = \sigma^2_A K + \sigma^2_c H + \sigma^2_{GxE} K \odot \Upsilon + \sigma^2_E I$$

Where $\odot$ is the Hadamard product and σ^2_{GxE} represents the variance of GxE effects. Υ matrix can be structured differently depending on the continuous or categorical nature of the environmental variable, i.e., for continuous variables it can be structured as a matrix of scaled similarities among individuals and can be modeled using an exponential decay, i.e., $p_{ij} = \exp(-\lambda|x_i - x_j|)$, while for categorical variables the matrix simply defines the individuals within the same environmental group. As for the previously described model, the sum of the variance components is constrained to 1, so that the proportion of total phenotypic variance, on a liability scale, explained by GxE effects, i.e., h^2_{GxE}, is derived diving σ^2_{GxE} by $1 + \sigma^2_\beta$. A posterior distribution for h^2_{GxE} parameter significantly greater than 0 indicates that the differential impact of additive genetic effects across different levels of environmental exposure has a statistically significant explanatory role for trait variability on the liability scale [132], considering all other explanatory components included in the model. Consequently, a null h^2_{GxE} parameter must not be interpreted as evidence for the absence of causal GxE effects. Instead, it suggests a negligible explanatory role for GxE effects in explaining trait variability.

2.1.5 Description of the Bayesian-LTHM framework

For the following explanations, an ACE model, i.e., $\Sigma = \sigma_A^2 K + \sigma_C^2 H + \sigma_E^2 I$, including B covariates will be considered. Due to the ascertainment scheme from sampled affected probands the sample is not representative of the population and most likely present cases with a higher frequency. LTMH can deal with this problem constraining each L_{ij} lower and upper bounds (a_{ij}, b_{ij}) according to the individual's observed phenotype Y_{ij} and a fixed threshold c which defines, above its value, the portion of the distribution equal to the cases prevalence in the population. The threshold c can be determined as the inverse of the cumulative distribution function evaluated at the trait's prevalence in the population [133], i.e., $\Phi(c)=P(L\geq c)$, where $L \sim N(0,1)$. Therefore, L_{ij} is bounded by $(-\infty, c)$ if the individual is a control, and, instead, bounded by $(c, +\infty)$ if the individual is a case, allowing to consider the unobserved liability scores as if they would have been sampled from the population L distribution. The joint probability density function (pdf) of the complete data $p(Y,L)$ can be decomposed into the marginal pdf of L and the conditional pdf of Y given that L has the support of (a, b):

$$p(Y, L) = p(Y|L)p(L) = p(L)I(a < L < b)$$

In this proposed method, a Bayesian framework is used to estimate a set of plausible values for parameters $\theta=(\sigma^2_A, \sigma^2_C, \sigma^2_E, h^2, c^2, e^2, b, \sigma^2_\beta, \tau^2_\beta)$, where the conditional sampling distribution, or likelihood, for the observed data Y and unobserved liabilities L, is defined by a truncated multivariate normal distribution, bounded in the range (a, b) depending on the observed phenotypes Y:

$$p(Y, L \mid \theta) = L \sim MVN(Xb, \Sigma)I(a < L < b)$$

However, due to nonrandom sampling, covariates distribution could not be representative of the target population distribution and b parameters require ascertainment bias correction. One method to correct ascertainment bias is to condition the likelihood on the proband's phenotypic information [134]. This approach is called "ascertainment-assumption free" (AAF) [134] as there is no need to explicitly model how ascertainment depends on phenotypes [135]. The sampling distribution for θ parameters adjusted from ascertainment bias is then defined as the following conditional likelihood:

$$p(Y^{NP}, L^{NP} \mid Y^P, L^P, \theta) = \frac{p(Y, L \mid \theta)}{p(Y^P, L^P \mid \theta)}$$

Where P denotes probands and NP non-probands. Probands P are assumed to be independent of each other. The numerator represents the likelihood function as previously described, while the denominator represents the likelihood that the proband is randomly picked from the population:

$$p(Y^P, L^P \mid \theta) = \prod_{i=1}^{F} (\exp\,(Y_i^P * \log\,(\frac{\mu_i}{1-\mu_i})) * 1 - \mu_i)$$

Where μ_i represents the probability that the liability score for a proband is higher than the threshold c, i.e., $\mu_i = P(Y_i^P=1)=P(L_i^P > c)=1- \Phi(c- X_i^P b)$. If all the probands are cases ($Y_i^P=1$, for each family i), then the likelihood simply reduces to:

$$p(Y^P, L^P \mid \theta) = \prod_{i=1}^{F} \mu_i$$

The posterior distributions $p(\theta \mid Y, L)$ can be then characterized using Bayes' theorem as:

$$p(\theta \mid Y, L) \propto p(Y^{NP}, L^{NP} \mid Y^P, L^P, \theta)p(\theta)$$

Where $p(\theta)$ represents the prior distribution specified for the parameters in θ.

2.1.6 Stan implementation and prior distributions specification

The computation of the Bayesian model relied on the program Stan [136] and the respective R interface package CmdStanR [137]. Stan makes use of the previously mentioned NUTS sampler [113], an extension of HMC methods [138], to draw samples from the parameters' posterior distributions. CmdStanR allows to improve speed efficiency with the support of between and within-chains multi-threading for parallelization. Between-chains parallelization allows to run MCMC chains in parallel on different cores (one for chain), while the within-chains parallelization allows, through the "reduce_sum" function, to partition the overall log-likelihood into arbitrary smaller partial log-likelihoods calculated in parallel on different cores (one for each partition specified by the user) [139]. Specifications for the prior distributions $p(\theta)$ will now be discussed. Since σ^2_A and σ^2_C are defined in the range $[0,1]$, a natural choice is to sample from a Beta distribution with shape hyperparameters α and β. Non-informative priors can be implemented fixing α and β hyperparameters to 1, e.g., $p(\sigma^2_A) \sim \text{Beta}(1,1)$, which is the equivalent of a uniform distribution bounded by 0 and 1. σ^2_E is specified as a transformed parameter, as for the constriction described above, and therefore does not need a prior distribution specification. Prior distribution for b parameters can be defined using a distribution with a support based on real values such as a normal distribution with customized hyperparameters μ and σ; a reasonably non-informative prior can be formulated as $p(b) \sim N(0,10)$. σ^2_E and σ^2_β are specified as transformed parameters, as for the formulas described above, and therefore do not need a prior distribution specification. h^2, c^2, e^2, and τ^2_β, which represent the parameters of interest for the interpretation of the results, can be treated in STAN as generated quantities or as transformed parameters using the formula described above. While the first option is computationally faster as its aim is simply to generate a sampled posterior distribution, the second allows the user to specify a more informative prior distribution, which is useful in case one is interested in adding previous knowledge gathered from past research studies. α and β hyperparameters can then be set to obtain the desired Beta distribution shape, e.g., $p(h^2) \sim \text{Beta}(\alpha, \beta)$. Moreover, one can rely on a Beta distribution re-parametrization to obtain α and β based on expected value $E(\theta)=\mu$ and precision parameter $\varphi > 0$. Larger φ values lead to smaller variance of θ, and φ can be set to obtain the desired variance using the formula $\frac{\mu(1-\mu)}{1+\varphi}$. Finally, given the chosen μ and φ parameters, shape hyperparameters α and β can be derived as $\alpha=\mu\varphi$ and $\beta=\varphi(1-\mu)$. Even though prior distributions are placed for transformed parameters, Jacobian adjustment of log-likelihood is not required since the transformations are linear, i.e., the sum of log of the absolute derivative of the determinant of the Jacobian matrix is a constant [139].

2.1.7 Bayesian-LTMH: STAN code implementation

Here, the reader can find a description for the STAN code [136] produced to implement Bayesian-LTMH (when one covariate is included in the model) following the previously illustrated ACE model.

1. The "data" chunk is used to define the data to be used along the code and which must be imported using the R interface.

```
data {
int<lower=0> N; //Number of subjects
matrix[N,N] kinship_matrix; //Kinship matrix
matrix[N,N] sharedenvironment_matrix_sibs; //Shared environment matrix
matrix[N,N] identity_matrix;  //Identity matrix
real threshold; //Normal distribution quantile which leaves to the right an area equal to the disease prevalence.
vector[N] lb; //Normal distribution lower bound based on the subject status (case-control)
vector[N] ub; //Normal distribution upper bound based on the subject status (case-control)
vector<lower=0,upper=1>[N] lb_ind; //Indicator of cases.
vector<lower=0,upper=1>[N] ub_ind; //Indicator of controls.
int<lower=0> NFAM; //Number of families.
int ni[NFAM]; //Number of subjects within the families.
int firstfam[NFAM]; //Placeholder for the first subject in the family.
int<lower=0> NCOV; //Number of covariates.
matrix[N,NCOV] X; //Covariate values.
vector[NCOV] SD; //Covariate’s standard deviations.
matrix[NFAM,NCOV] XP; //Covariate values for the family’s proband.
vector[NFAM] YP; //Status (case-control) for the family’s proband.
int<lower=1> fam[NFAM]; //Sequence for number of families.
int<lower=1> grainsize; //Granularity of within-chain parallelization (default=1).
}
```

2. In the "parameters" chunk the unknown parameters to be drawn are defined.

```
//Here, initial parameters are defined.
parameters {
vector[NCOV] betapam; //Fixed-effect parameter.
real <lower=0, upper= 1> sharedenvironment_sibs; //Shared environment effect variance.
real <lower=0, upper= 1-sharedenvironment_sibs> heritability; //Additive genetic effects variance.
vector<lower=0,upper=1>[N] u; //Latent parameters for the liability scores.
}
```

3. In the “transformed parameters” chunk the parameters derived from sum constraints are defined.

```
//Individual environment effects variance is derived from sum constraint, so that the sum is equal to 1.
transformed parameters{
real <lower=0, upper= 1> individualenvironment = 1-heritability-sharedenvironment_sibs;
}
```

4. The “generated quantities” chunk is useful to obtain a posterior distribution for the quantities of interest derived in last place as a derivation of the parameters in the model. Defining these quantities in this chunk, instead of “transformed parameters” greatly improves the computational efficiency.

```
generated quantities{
//Variance explained by the fixed-effect.
real sigmaCOV=variance(X*betapam);

//Marginal parameters posterior distributions.
//Heritability:
real <lower=0, upper= 1> heritability_marginal = heritability/(1+sigmaCOV);
//Proportion of variance explained by shared environment effects:
real <lower=0, upper= 1> sharedenvironment_sibs_marginal = sharedenvironment_sibs/(1+sigmaCOV);
//Proportion of variance explained by individual environment effects:
real <lower=0, upper= 1> individualenvironment_marginal = individualenvironment/(1+sigmaCOV);
//Proportion of variance explained by the fixed-effect:
real <lower=0, upper= 1> tauCOV_marginal = sigmaCOV/(1+sigmaCOV);

//Unstandardized fixed-effect:
real beta_COV = betapam[1] / SD[1];
}
```

5. The log-likelihood function is defined assuming liabilities to be distributed as truncated multivariate normal distribution.

```
functions {
//The partial_sum function is necessary to add computational efficiency using parallel cores.
real partial_sum(int[] fam_slice, int start, int end, matrix X, matrix XP, vector YP, int[] ni, int[] firstfam, vector lb,
vector ub, vector lb_ind, vector ub_ind, vector u, int NFAM, real heritability, real sharedenvironment_sibs, real
individualenvironment, matrix kinship_matrix, matrix sharedenvironment_matrix_sibs, matrix identity_matrix, vector
betapam, int NCOV, real threshold) {

//Likelihood initialization.
        real lik=0;

//Within this loop, each family is selected once at a time to calculate the log-likelihood.
         for (k in start:end) {
```

```
//Each element specific to the family is selected.
        int nik =ni[k];
        int firstfamk =firstfam[k];
        row_vector[NCOV] XPk =XP[k,];
        real YPk =YP[k];
        matrix[nik,NCOV] Xk=block(X, firstfamk ,1,nik,NCOV);
        vector[nik] lbk=segment(lb, firstfamk ,nik);
        vector[nik] ubk=segment(ub, firstfamk ,nik);
        vector[nik] lb_indk=segment(lb_ind, firstfamk ,nik);
        vector[nik] ub_indk=segment(ub_ind, firstfamk ,nik);
        vector[nik] uk=segment(u, firstfamk ,nik);
        vector[nik] mu = Xk*betapam;

//Ascertainment bias adjustment
        real XB = XPk*betapam;
        real alpha=YPk*log((1-normal_cdf(threshold - XB, 0, 1))/(1-(1-normal_cdf(threshold - XB, 0, 1)))) - log(1/(1-
        (1-normal_cdf(threshold - XB, 0, 1))));

//Covariance matrix definition.
        matrix [nik,nik] Sigmakc=cholesky_decompose(block(kinship_matrix, firstfamk, firstfamk, nik, nik)*heritability
        + block(sharedenvironment_matrix_sibs, firstfamk, firstfamk, nik, nik)*sharedenvironment_sibs
        + block(identity_matrix, firstfamk, firstfamk, nik, nik)*(individualenvironment));

//Calculation of the truncated multivariate normal distribution log-likelihood based on the parameters drawn.
          vector[nik] z;
            real prob = 0;
             for ( m in 1:nik ) {
               if ( lb_indk[m] == 0 && ub_indk[m] == 0 )  z[m] = inv_Phi(uk[m]);
                 else {
                   int km1 = m - 1;
                   real v;
                   real z_star;
                   real logd;
                   row_vector [2] log_ustar = [negative_infinity(), 0];
                   real constrain = mu[m] + ((m > 1) ? Sigmakc[m, 1:km1] * head(z, km1) : 0);
                   if ( lb_indk[m] == 1 ) log_ustar[1] = normal_lcdf( ( lbk[m] - constrain ) / Sigmakc[m, m] | 0.0, 1.0 );
                   if ( ub_indk[m] == 1 ) log_ustar[2] = normal_lcdf( ( ubk[m] - constrain ) / Sigmakc[m, m] | 0.0,1.0);
                   logd  = log_diff_exp(log_ustar[2], log_ustar[1]);
                   v  = exp( log_sum_exp( log_ustar[1], log(uk[m]) + logd ) );
                   z[m] = inv_Phi(v);
                   prob += logd;
                 }
               }
                 lik += prob-alpha;
        }

          return  lik;
        }
        }
```

6. The parameters' prior distributions and the target distribution, as previously defined in the function "partial_sum", are specified in the "model" chunk.

```
model {
//For the fixed-effect, a non-informative normal distribution is selected.
betapam~ normal(0,10);
//For heritability, a non-informative beta distribution is selected.
heritability~ beta(1,1);
//For shared-environment effects variance, a non-informative beta distribution is selected.
sharedenvironment_sibs~ beta(1,1);

//Target truncated multivariate normal distribution. Reduce_sum function is needed to operate
parallelization and increase computational efficiency.
target += reduce_sum(partial_sum,fam , grainsize,X, XP, YP,ni, firstfam, lb, ub, lb_ind, ub_ind, u,
NFAM,heritability,sharedenvironment_sibs, individualenvironment,
kinship_matrix,sharedenvironment_matrix_sibs,identity_matrix, betapam,NCOV,threshold);
  }
```

2.1.8 Simulation studies

To assess the ability of the proposed Bayesian-LTMH to recover the true parameters, simulations were performed under different scenarios. The aim was to evaluate the accuracy and precision of the posterior distribution for the parameters of interest according to the model specification, pedigree structure and trait's prevalence. Therefore, posterior distribution uncertainty relative to the prior knowledge, i.e., standard deviation (SD) and lack of bias were evaluated. To answer these questions, different scenarios were simulated according to i) the pedigree structure, sampling 500 nuclear families or 150 three-generations families from affected probands, ii) trait's prevalence, i.e., 0.05 and 0.005, and iii) the model specification and the effects used to simulate liability scores, which include different combinations of additive genetic effects (A), shared environment effects (C), dominant genetics effects (D), individual environment effects (E), as well as the effect of a single-nucleotide polymorphism (SNP) covariate β_{SNP}. The detailed steps were the following:

1. First, 150,000 nuclear families, with parents having 1, 2, 3, or 4 sons/daughters with probabilities 0.2, 0.3, 0.3, 0.2 were randomly simulated. Sex was always assigned with probability 0.5, and one of the sons/daughters was randomly chosen to represent the proband. In an alternative scenario 40000 extended pedigrees up to the third generation were randomly simulated. The first generation was made of a founders' couple having 2, 3 or 4 sons/daughters with probabilities 0.4, 0.4 and 0.2 which themselves had 1, 2, 3 or 4 sons/daughter with the probabilities 0.2, 0.3, 0.3 and 0.2. Among the second and the third generation, an individual was randomly chosen to represent the proband.
2. Within each family, liabilities were simulated as random draws from a multivariate normal distribution with covariance matrix equal to the sum of the effects specified depending on the scenario and mean equal to 0 or equal to $X\beta_{SNP}$ depending on SNP covariate being included in the model. The performance of Bayesian-LTMH was evaluated fitting: 1) AE model, when liabilities were simulated fixing $h^2 = 0.4$, null c^2, no covariates included; 2) ACE model, modeling c^2_{Sibs}, when liabilities were simulated fixing $h^2 = 0.4$, $c^2_{Sibs} = 0.2$, no covariates

included; 3) ACE model, modeling c^2_{Sibs} and $c^2_{Mother\text{-}Offspring}$, when liabilities were simulated fixing $h^2 = 0.4$, $c^2_{Sibs} = 0.2$, $c^2_{Mother\text{-}Offspring} = 0.1$, no covariates included. 4) ACE model as in 2), when liabilities were simulated fixing $h^2 = 0.4$, $c^2_{Sibs} = 0.2$, $d^2=0.2$, no covariates included, to quantify the potential bias in h^2 and c^2 parameters posterior distributions when dominant genetic effects are present but not accounted in the model. 5) ACE model as in 2) but including a SNP as covariate, when liabilities were simulated fixing $h^2 = 0.4$, $c^2_{Sibs} = 0.2$, and SNP effect β_{SNP} explaining 1% of total phenotypic variance, i.e., $h^2_{SNP} = 0.01$. Founder genotypes for each family were generated from a binomial distribution with two trials and the Minor Allele Frequency (MAF) as success probability, which was fixed to 0.2. Non-founder genotypes were consequently obtained following Mendelian transmission. To obtain $h^2_{SNP} = 0.01$, β_{SNP} was fixed to 0.178 following the equation [96]:

$$h^2{}_{SNP} = \frac{2 * \beta_{SNP}{}^2 * MAF * (1 - MAF)}{1 + 2 * \beta_{SNP}{}^2 * MAF * (1 - MAF)}$$

Once liabilities were generated, individuals were considered as cases if their liabilities were larger than a threshold c, which was chosen to maintain the desired cases prevalence. Depending on the scenario, prevalence was fixed as 0.05 or 0.005. Finally, 500 nuclear families and 150 three generations families were randomly sampled between families with an affected proband, for an expected sample size of ≈ 2400 individuals.

3. Once the ascertained family-based sample was obtained, the Bayesian-LTMH specified according to the scenario was implemented using NUTS to draw samples from the posterior distribution, setting two chains with 1000 warmup iterations and 1000 sampling iterations. Prior distributions were fixed as non-informative Beta distribution, i.e., Beta(1,1), for h^2, c^2_{Sibs} and $c^2_{Mother\text{-}Offspring}$ parameters, and as non-informative normal distribution. i.e., N(0,10), for β_{SNP} parameter.
4. The points 1-3 were repeated 200 times for each scenario. From the obtained parameters' sampled posterior distributions, different descriptive statistics useful to evaluate the performance of Bayesian-LTMH were calculated. The median of the posterior distribution was considered as a point estimate. To evaluate the accuracy of parameters posterior distributions across all 200 simulations, it was calculated 1) the median of all point estimates and 2) the bias as the difference from the respective true parameter value. To evaluate the precision, it was calculated 3) the SD of all point estimates and 4) the median of all posterior distributions' SDs. Moreover, it was calculated 5) the root mean square error (RMSE) as a measure to compare the quality of the posterior distribution, both in terms of accuracy and precision, between scenarios. RMSE is defined as the square root of the mean square difference between the point estimates and the respective true value, i.e., $\sqrt{E[(\widehat{\theta^2} - \theta^2)^2]}$. Finally, it was calculated 6) the coverage as the number of times the 95% Highest Posterior Density Credibility Intervals (HPD CIs) contained the true parameter.

The computational time to fit AE and ACE models using LTHM under the Bayesian framework and under the EM-algorithm approach was compared, considering a sample of 150 three-generations families ascertained from a proband where trait prevalence was equal to 0.005.

2.2 Implementing Bayesian-LTMH on the Sardinian extended families

2.2.1 Sardinian Families Ascertainment

The sample used in this work was retrieved from a register of MS cases, diagnosed according to Poser's criteria [28], established in Sardinia's Nuoro province in 1995. Whenever possible, patients were examined by the neurologists at the Neurology Department of the Nuoro Hospital. Otherwise, clinical records were obtained and reviewed by the previous neurologists. During the examination, the neurologists filled the clinical record of the patient, comprising the MS disease course. From this case register, 89 MS-affected probands were sampled, without any selection in favor of MS patients with a possible family history. Using the genealogical questionnaires filled in by the affected proband and the municipal registries, it was possible to reconstruct their genealogical tree. In some cases, MS probands resulted distantly related through a common ancestor, leading to a final sample comprising 24 extended families [97]. Examples of extended families are reported in **Figure 13**. In this analysis, probands' parents, siblings, spouses, uncles/aunts, first-degree cousins, nieces/nephews, and grandparents were included ,while more distant relatives were excluded to avoid MS misclassifications. Non-affected relatives included in the final analysis were at least 20 years old at the day of the questionnaire compilation. Thus, a total of 790 subjects were analyzed, comprising 118 MS cases and 672 healthy controls.

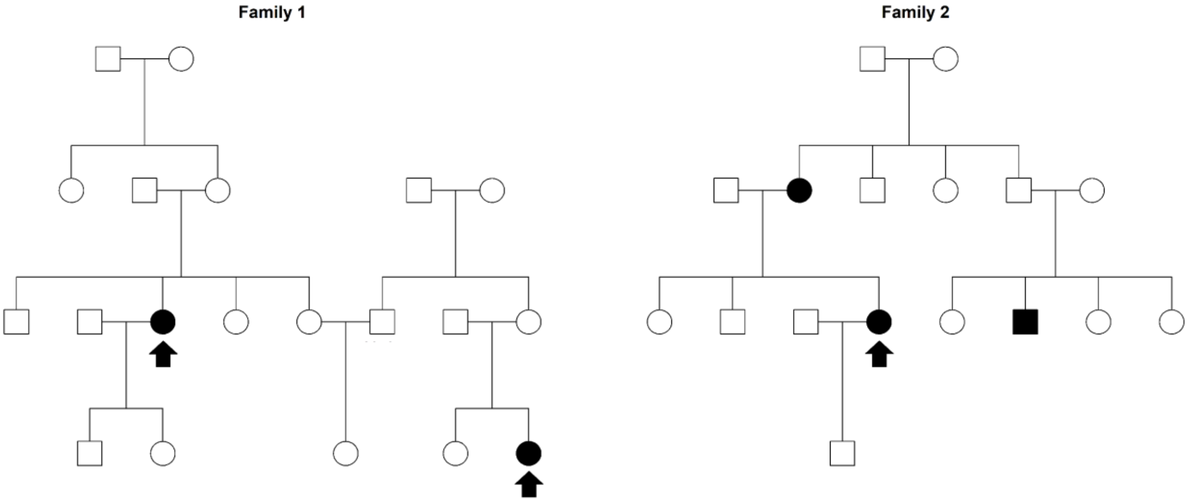


Figure 13. Examples of Sardinian extended families. Multiple Sclerosis cases are reported in black, while the arrow denotes a proband.

2.2.2 Model specification

To estimate MS heritability implementing the Bayesian-LTMH method, making use of the Sardinian family-based sample, a liability threshold model was specified. To define the fixed threshold (c), which depends on MS prevalence in the population [133], the work by Montomoli et al. [140] was used, which estimated MS crude prevalence in Nuoro province as 157 per 100,000 inhabitants on December 31, 1998 (around the years of questionnaires compilation as reported above). To adjust for potential confounding, the following covariates were included in the model,: (i) sex, as the female-to-male MS prevalence ratio in the Nuoro province was reported to be 2:1 [9]; (ii) categorized year of birth (<1946 or ≥1946) as a proxy for the individuals' different early environmental exposures (comprising the beginning of malaria eradication program). L was assumed to be distributed following a multivariate normal distribution, i.e., $L \sim MVN(Xb, \Sigma)$, where X denotes a matrix for standardized covariates, i.e., sex and categorized year of birth (YR), b represents the respective vector of fixed effects parameters, i.e., β_{SEX} and β_{YR}, and Σ denotes a covariance matrix. The AAF approach was applied to correct fixed effects' ascertainment bias. Since multiple distant related probands could be present within a single family i, it was considered, for simplicity, a single fictitious proband with covariates values equal to the mean of the actual probands' sex and categorized year of birth within the family i, i.e., $X_{ij}^{P} = \frac{1}{M}\sum_{m=1}^{M} X_{ijm}^{P}$, where $m = 1, \ldots, M$ were the respective probands in family i and $j=1,2$ were the respective covariates. Liability scores for the family members were then assumed to be normally distributed and correlated as follows:

$$L_i \sim MVN(X_1\beta_{Sex} + X_2\beta_{YR}, \Sigma)$$

The standard polygenic additive model [67,114] was applied, assuming null epistatic and gene–environment (G×E) effects, defining Σ as follows (ACE model):

$$\Sigma = h^2K + c^2_{Sibs}H_1 + c^2_{Mother-Offspring}H_2 + c^2_{Father-Offspring}H_3 + c^2_{Spouses}H_4 + e^2I,$$

where parameters were defined as the proportion of MS liability variability explained by:

- h^2, additive genetic effects, with K being the kinship matrix multiplied by two,
- c^2_{Sibs}, effects due the environment shared between siblings (which also allow to adjust for dominant genetic effects), with H_1 being the correlation matrix with values equal to 1 between siblings,
- $c^2_{Mother-Offspring}$, effects of environment shared between the mother and the offspring, which may include maternal effects as highlighted in [141,142], with H_2 being the correlation matrix with values equal to 1 between mother and offspring,
- $c^2_{Father-Offspring}$, effects of environment shared by the father and the offspring, with H_3 being the correlation matrix with values equal to 1 between father and offspring,
- $c^2_{Spouses}$, effects of environment shared between spouses, with H_4 being the correlation matrix with values equal to 1 between spouses,
- e^2, individual environmental effects, with I being the respective identity matrix. To avoid identifiability problems [129], e^2 was derived as the complementary to 1 considering the sum of the other parameters.

The proportion of MS liability variance explained by total shared environment effects, i.e., c^2_{Total}, was then defined as the sum of c^2_{Sibs}, $c^2_{Mother-Offspring}$, $c^2_{Father-Offspring}$, and $c^2_{Spouses}$ components. Modeling c^2_{Total} allows avoiding an inflation in h^2 due to common environmental influences [82,119,143]. β_{SEX} and β_{YR} allow quantifying the liability increase/decrease and the proportion of MS liability variability jointly explained by both covariates, i.e., $\tau^2_{\beta\text{SEX,YR}} = \text{var}(\text{Xb})$ [116]. This latter term can be decomposed, following [144], into

$$\tau^2_{\beta\text{SEX,YR}} = \tau^2_{\beta\text{SEX}} + \tau^2_{\beta\text{YR}} + 2\text{cov}_{\beta_{\text{SEX,YR}}},$$

from which was derived the proportion of MS variability marginally explained by (i) sex, $\tau^2_{\beta\text{SEX}}$, (ii) categorized year of birth, $\tau^2_{\beta\text{YR}}$, and (iii) their covariance component, $2\text{cov}_{\beta\text{SEX,YR}}$. As described in [116], $\tau^2_{\beta\text{SEX,YR}}$ was considered as part of the total phenotypic variance to obtain marginal posterior distributions. Posterior distributions for β_{SEX} and β_{YR} parameters were unstandardized dividing the values by the variables' SD. Using the above-specified model, two separate analyses were conducted. In the first, the whole sample was considered, aiming to quantify the overall contribution of genetic and environmental variability in explaining MS susceptibility variability considering individuals born before and after World War II and malaria eradication. The explanatory role of G×E effects, between additive genetics and categorized year of birth, was also assessed in a separate model [131,132]. In the second, the sample was stratified based on the categorized year of birth; the rationale was to evaluate the explanatory influence of genetic and environmental factors on subgroups of individuals with more similar early environmental exposures linked to the year of birth. To better reflect the MS prevalence in these two groups, the work of Montomoli et al. [140] was used to set MS prevalence as 103 per 100,000 inhabitants for the individuals born before 1946, and as 176 per 100,000 inhabitants for the individuals born on/after 1946. Only for the analysis on individuals born on/after 1946 it was possible to include the exact year of birth as a continuous covariate in the model to investigate the temporal change in MS liability. Given the lack of previous results for h^2 estimation in the Sardinian population, non-informative prior distributions were selected for all parameters, i.e., $\text{Beta}(1,1)$ for variance components, and $\text{N}(0,10)$ for β_{SEX} and β_{YR} parameters. Following the suggestions of STAN developers, to obtain the sampled parameters' posterior distributions four chains with 5000 warmup iterations and 5000 sampling iterations were run, for a total of 20,000 sampling iterations, and convergence of the four chains to the same posterior distribution was assessed visually using trace plots and inspecting diagnostic summaries as provided by STAN software. All analyses were performed using RStudio and Stan softwares [136–138].

3 RESULTS AND DISCUSSION

3.1 Simulations studies

Posterior distributions for each parameter were obtained by sampling via the NUTS sampler implemented in the program STAN [136]. The performance was evaluated in terms of accuracy, precision, and coverage; **Table 1** reports the descriptive statistics for the parameters posterior distributions obtained within each simulated scenario, while **Figure 14** reports the corresponding box plots with a red line indicating the true parameter value. No divergences or other diagnostic problems were encountered during NUTS sampling. Considering all the scenarios, point estimates for all parameters were generally close to the true value. Therefore, accurate h^2 were obtained in presence of confounders such as shared environmental effects. It can be observed that the RMSE and posterior distribution SD of the estimator across different scenarios showed an increase with i) a lower trait prevalence, or/and ii) increasing the number of variance components in the model, or/and iii) using three-generations families. The latter result can be explained due to decreasing genetic relatedness among distant relatives within a family, such as grandparents-grandchildren or nephews/nieces-uncles/aunts, which led to a lower statistical power compared to the scenario with nuclear families and same sample size. Ascertainment bias was correctly adjusted for β_{SNP} when a SNP covariate was included in the ACE model. A slight downward bias for h^2 parameter was observed when an additional shared environment effect variance component, i.e., $c^2_{Mother\text{-}Offspring}$, was included in the ACE model; this bias was higher when the prevalence of the trait was equal to 0.005 and using three-generations families. When dominance genetic effects variance d^2=0.2 was included in liabilities simulation but not accounted for in the ACE model, the medians of c^2_{Sibs} posterior distributions obtained were, as expected, inflated by a factor corresponding to $0.25d^2=0.05$. However, this adjustment allowed to obtain accurate h^2 posterior distributions, avoiding the inflation from both c^2_{Sibs} and d^2 confounding. Finally, HPD CIs coverage was generally near to 95% in each scenario. Regarding computational efficiency, STAN employed 358.7 seconds to fit an AE model running one chain with 1000 warmup iterations and 1000 sampling iterations, without requiring multi-threading within-chain parallelization, on a sample of 150 three-generations families. Considering the same sample and number of fixed iterations, STAN employed 401.8 seconds to fit an ACE model including a parameter for c^2. The computational time dropped, respectively, to 124.8 and 140.1 seconds when 10 threads were set for within-chain parallelization. Instead, considering the same sample and the same models, the EM-based approach took more than one hour to proceed with a second iteration even after setting 100 threads for parallelization, therefore highlighting the dramatic improvement in speed using the Bayesian framework.

Table 1. Descriptive statistics for the sampled posterior distributions obtained fitting Bayesian liability threshold model on the 200 simulated datasets within each different scenario.

Pedigree*	Trait Prevalence	Parameter	Point Estimate Median (SD)**	Bias	SD°	RMSE^	Coverage (95% CI)
AE model, true h^2 = 0.4							
Nuclear	0.05	h^2	0.393 (0.046)	-0.007	0.045	0.047	0.94
Three-generations			0.399 (0.061)	-0.001	0.063	0.061	0.95
Nuclear	0.005		0.398 (0.056)	-0.002	0.055	0.057	0.94
Three-generations			0.385 (0.084)	-0.015	0.082	0.087	0.94
ACE model true h^2 = 0.4, true c^2_{Sibs} = 0.2							
Nuclear	0.05	h^2	0.400 (0.054)	0.000	0.054	0.054	0.93
		c^2_{Sibs}	0.199 (0.036)	-0.002	0.038	0.036	0.96
Three-generations		h^2	0.399 (0.074)	-0.001	0.073	0.074	0.96
		c^2_{Sibs}	0.197 (0.050)	-0.003	0.051	0.050	0.95
Nuclear	0.005	h^2	0.387 (0.071)	-0.013	0.069	0.072	0.95
		c^2_{Sibs}	0.208 (0.045)	0.008	0.044	0.046	0.97
Three-generations		h^2	0.379 (0.096)	-0.021	0.099	0.098	0.92
		c^2_{Sibs}	0.206 (0.059)	0.006	0.067	0.060	0.97
ACE model, true h^2 = 0.4, true c^2_{Sibs} = 0.2, true β_{SNP}=0.178, true h^2_{SNP} = 0.01							
Nuclear	0.05	h^2	0.396 (0.052)	-0.004	0.053	0.052	0.95
		c^2_{Sibs}	0.200 (0.040)	0.000	0.038	0.040	0.95
		β_{SNP}	0.180 (0.062)	0.002	0.067	0.062	0.98
		h^2_{SNP}	0.012 (0.008)	0.002	0.009	0.008	0.98
Three-generations		h^2	0.388 (0.079)	-0.012	0.069	0.080	0.92
		c^2_{Sibs}	0.209 (0.047)	0.009	0.044	0.047	0.94
		β_{SNP}	0.188 (0.091)	0.010	0.096	0.090	0.96
		h^2_{SNP}	0.013 (0.012)	0.003	0.014	0.013	1.00
Nuclear	0.005	h^2	0.393 (0.071)	-0.007	0.073	0.072	0.96
		c^2_{Sibs}	0.202 (0.050)	0.002	0.051	0.050	0.97
		β_{SNP}	0.167 (0.074)	-0.011	0.071	0.074	0.93
		h^2_{SNP}	0.009 (0.009)	-0.001	0.008	0.009	0.93
Three-generations		h^2	0.378 (0.108)	-0.022	0.096	0.110	0.90
		c^2_{Sibs}	0.195 (0.066)	-0.005	0.064	0.066	0.92
		β_{SNP}	0.163 (0.131)	-0.015	0.119	0.130	0.93
		h^2_{SNP}	0.010 (0.017)	0.000	0.014	0.018	0.97
ACE model, true h^2 = 0.4, true c^2_{Sibs} = 0.2, true $c^2_{Mother\text{-}Offspring}$ = 0.1							
Nuclear	0.05	h^2	0.382 (0.073)	-0.018	0.073	0.075	0.94
		c^2_{Sibs}	0.202 (0.046)	0.002	0.044	0.047	0.94
		$c^2_{Mother\text{-}Offspring}$	0.102 (0.045)	0.002	0.048	0.046	0.95
Three-generations		h^2	0.385 (0.085)	-0.015	0.087	0.087	0.93
		c^2_{Sibs}	0.214 (0.052)	0.014	0.054	0.054	0.96
		$c^2_{Mother\text{-}Offspring}$	0.102 (0.046)	0.002	0.054	0.046	0.96

Nuclear	0.005	h^2	0.370 (0.091)	-0.030	0.091	0.098	0.93
		c^2_{Sibs}	0.218 (0.053)	0.018	0.053	0.056	0.93
		$c^2_{Mother\text{-}Offspring}$	0.103 (0.052)	0.003	0.056	0.053	0.95
Three-generations		h^2	0.353 (0.107)	-0.047	0.112	0.116	0.93
		c^2_{Sibs}	0.218 (0.061)	0.018	0.069	0.065	0.97
		$c^2_{Mother\text{-}Offspring}$	0.104 (0.045)	0.004	0.069	0.046	0.99
ACE model, true h^2 = 0.4, true c^2_{Sibs} = 0.2 and true d^2 = 0.2							
Nuclear	0.05	h^2	0.399 (0.056)	0.001	0.054	0.056	0.94
		$c^2_{Sibs}+0.25d^2$	0.256 (0.038)	0.006	0.037	0.038	0.97
Three-generations		h^2	0.404 (0.075)	0.004	0.073	0.075	0.96
		$c^2_{Sibs}+0.25d^2$	0.256 (0.052)	0.006	0.051	0.052	0.96
Nuclear	0.005	h^2	0.392 (0.074)	-0.008	0.069	0.074	0.91
		$c^2_{Sibs}+0.25d^2$	0.253 (0.048)	0.003	0.044	0.048	0.92
Three-generations		h^2	0.376 (0.104)	-0.024	0.099	0.107	0.94
		$c^2_{Sibs}+0.25d^2$	0.255 (0.067)	0.005	0.065	0.067	0.93

* 500 nuclear families or 150 three-generations families were obtained sampling affected probands

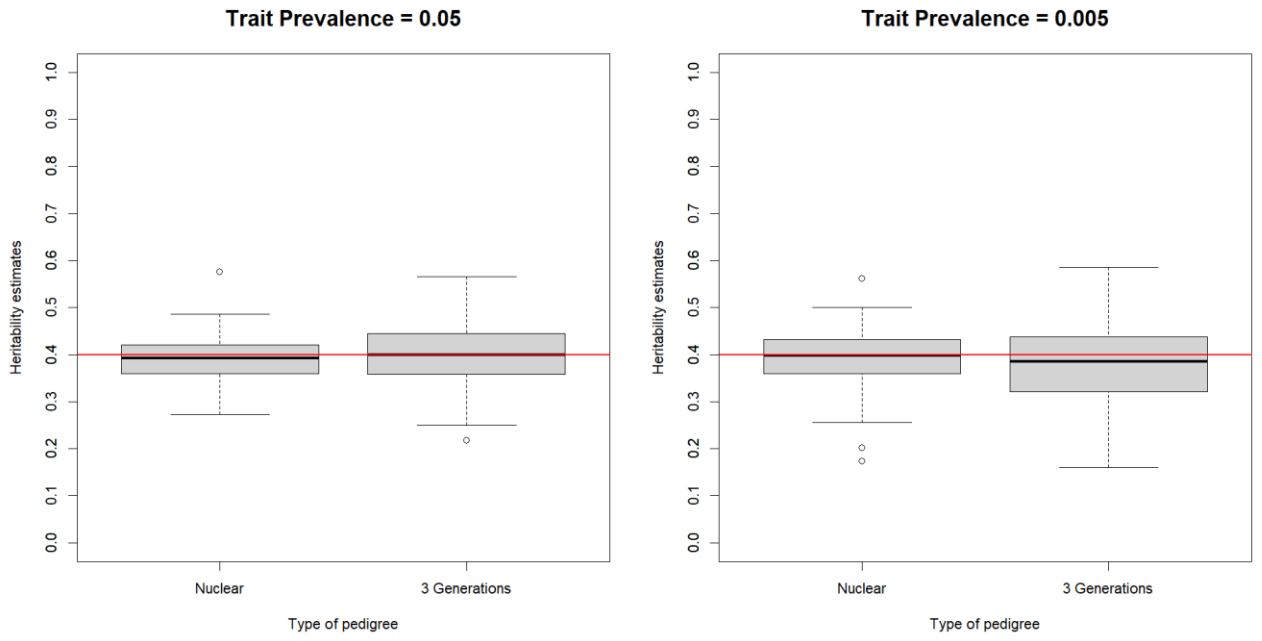
** The point estimate is represented by the median of the posterior distribution

° Median of all posterior distributions' standard deviations

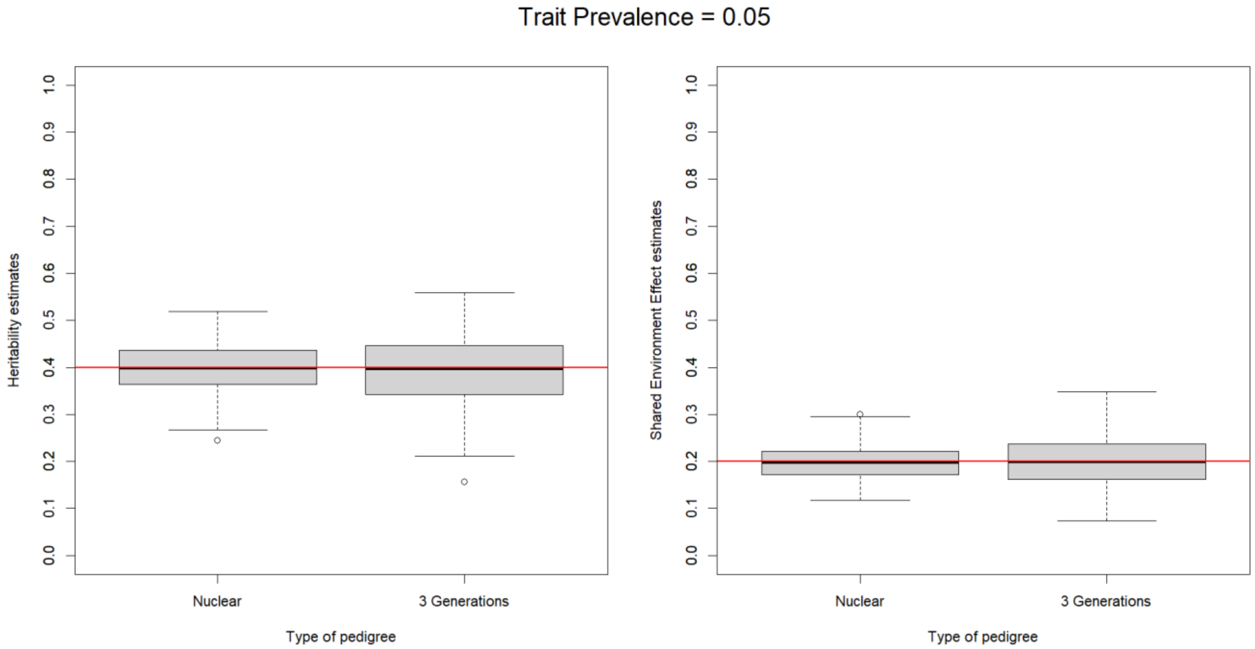
^ Root mean square error

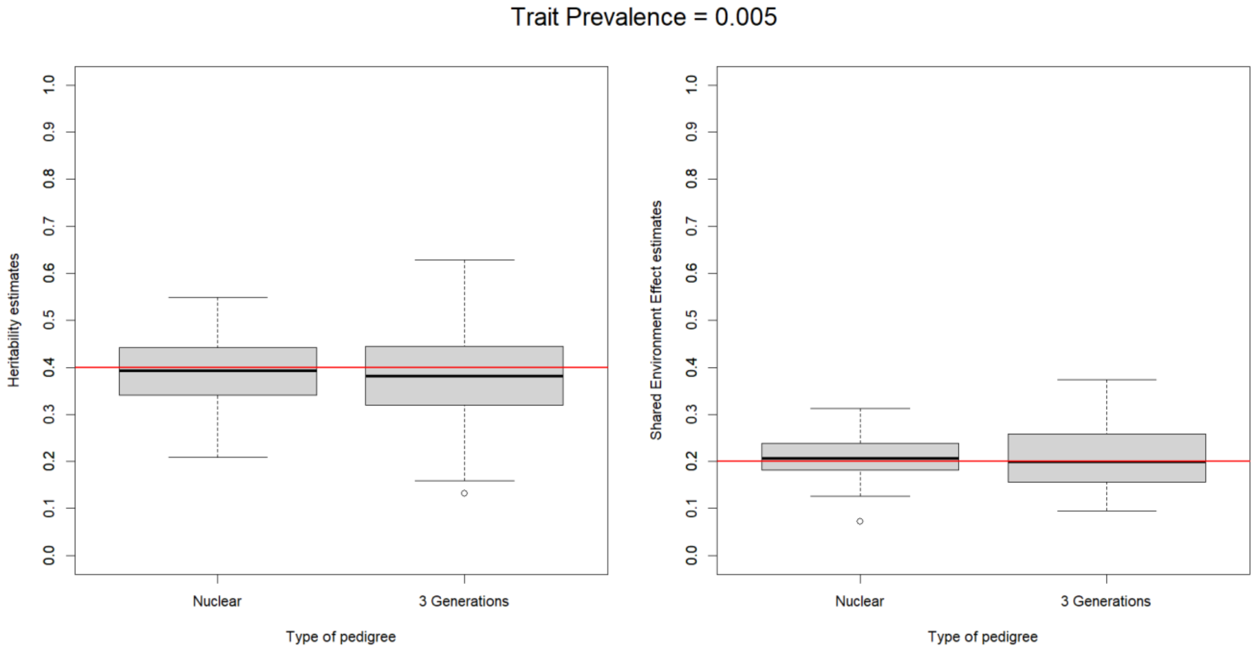
Figure 14. Box plots for the sampled posterior distributions obtained fitting Bayesian liability threshold model on the 200 simulated datasets within each different scenario.

1) AE model, true h^2 = 0.4



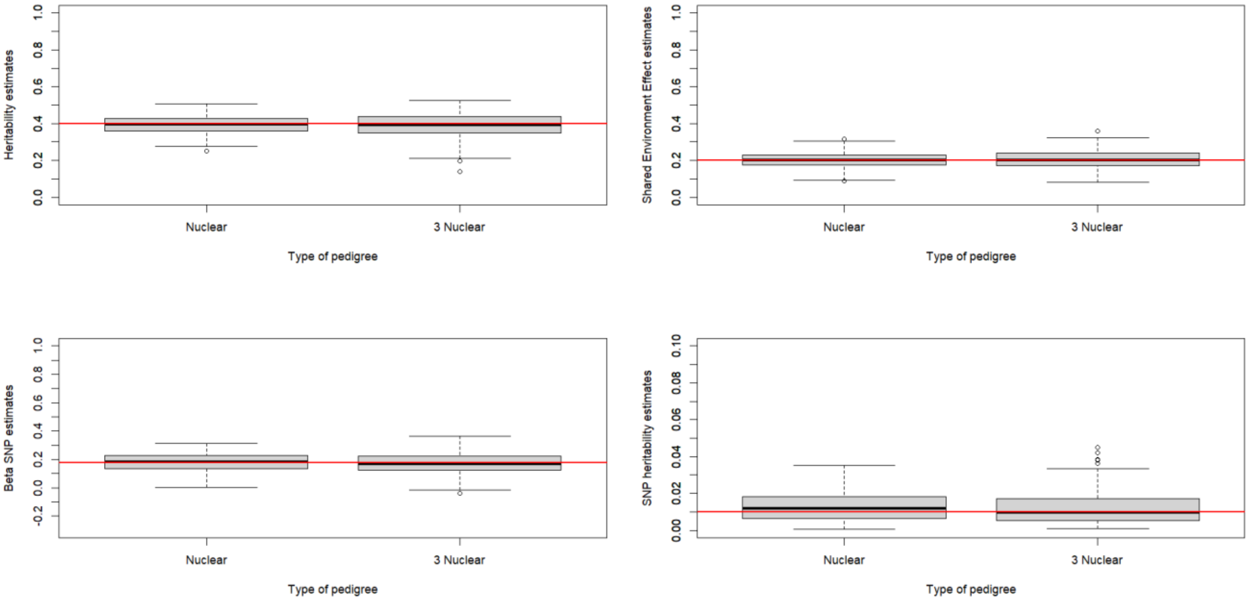
2) ACE model, true $h^2 = 0.4$, true $c^2_{Sibs} = 0.2$



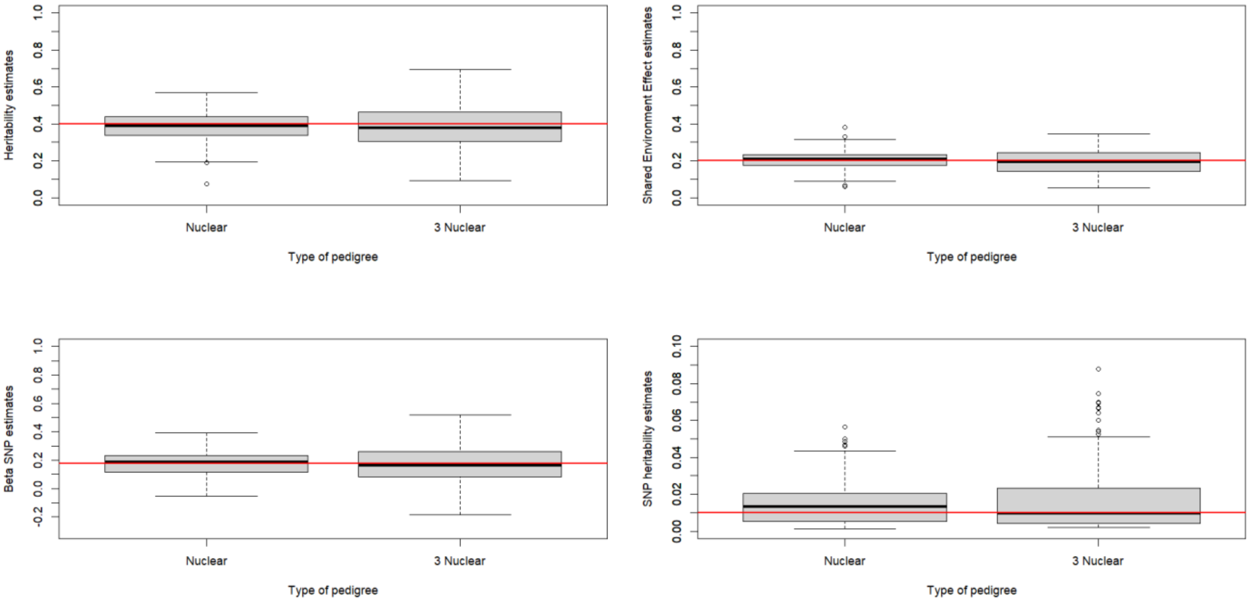


3) ACE model, true h^2 = 0.4, true c^2_{sibs} = 0.2, true β_{SNP}=0.178, true h^2_{SNP} = 0.01

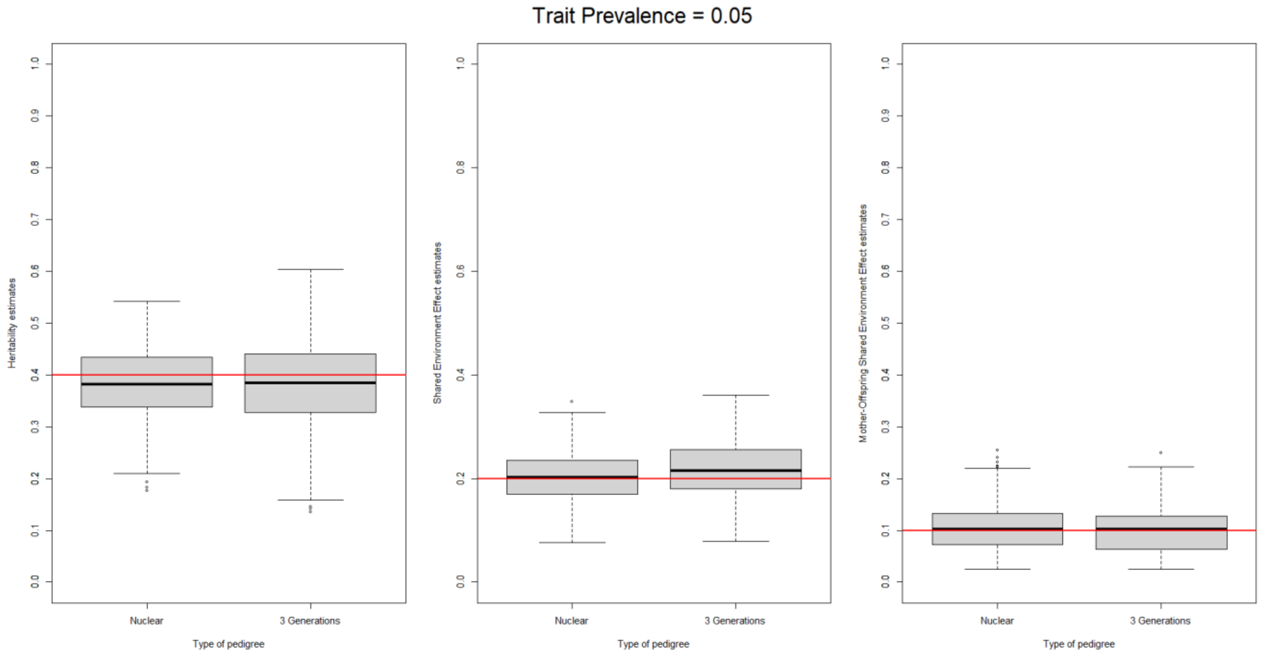
Trait Prevalence = 0.05

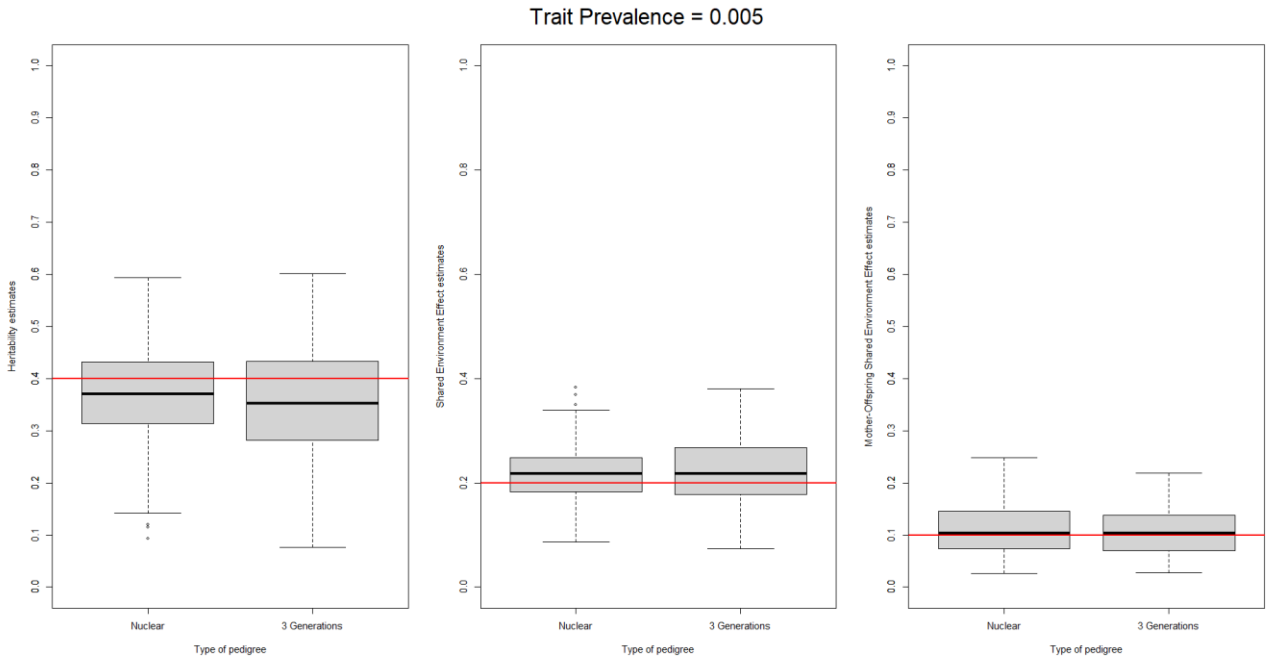


Trait Prevalence = 0.005



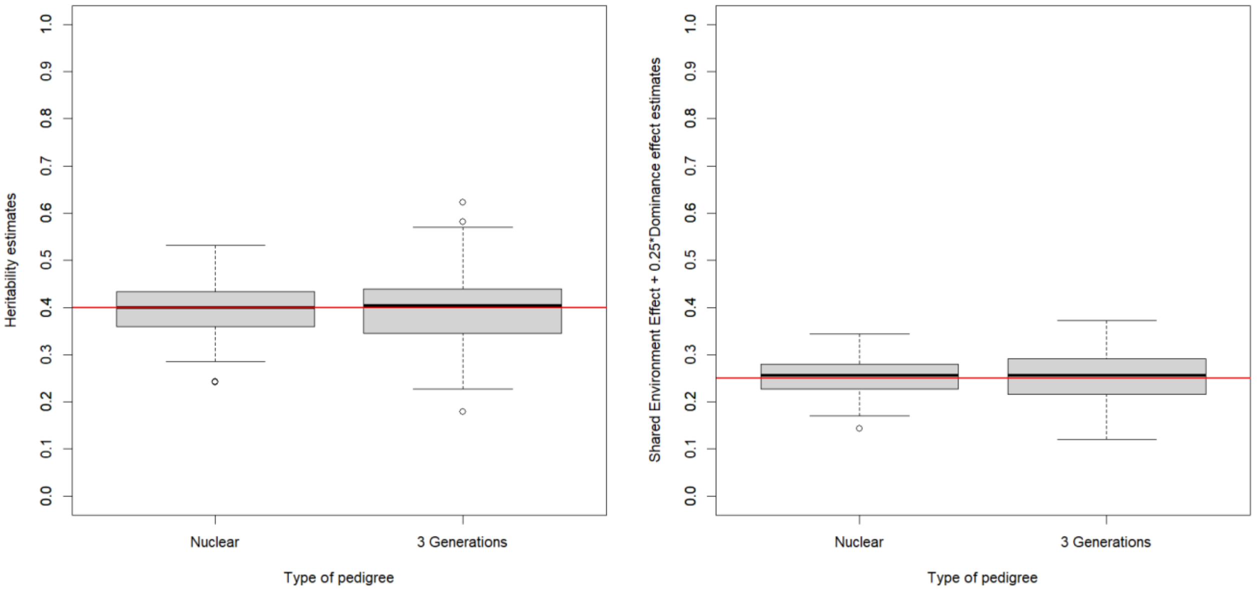
4) ACE model, true $h^2 = 0.4$, true $c^2_{Sibs} = 0.2$, true $c^2_{Mother\text{-}Offspring} = 0.1$



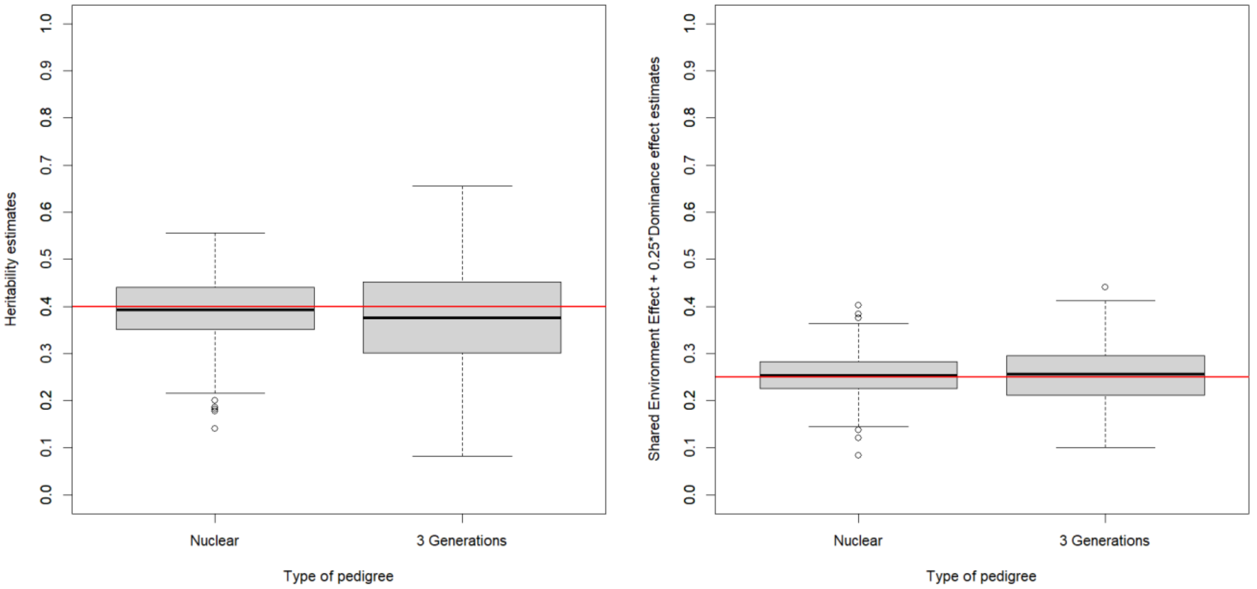


5) ACE model, true $h^2 = 0.4$, true $c^2_{Sibs} = 0.2$ and true $d^2 = 0.2$

Trait Prevalence = 0.05



Trait Prevalence = 0.005



3.2 Bayesian-LTMH application on the Sardinian extended families

The Bayesian-LTMH was implemented on the sample of 24 extended families from Sardinia's Nuoro province ascertained from MS affected probands, aiming to investigate the sources of MS susceptibility variability, among environmental and additive genetic effects, in the Sardinian individuals born across the 20th century.

3.2.1 Sample Description

The analyzed 24 Sardinian families each comprised 7 to 93 subjects (median = 26 subjects) and 1 to 16 MS cases (median = 3 MS cases), for a total of 790 subjects of which 118 were MS cases (15%) and 672 healthy related controls (85%). A total of 302 individuals (38%) were born on/after 1946. Among the 118 MS cases, 76 were females (64%) and 42 males (36%), which gave a female/male ratio equal to 1.81. Descriptive statistics are reported in **Table 2**.

Table 2. Descriptive statistics for the 24 Sardinian extended families.

Family	Individuals N (%) [1]	Probands* N	Females N (%) [2]	MS Cases N (%) [2]
1	65 (8%)	6	37 (57%)	6 (9%)
2	35 (4%)	4	20 (57%)	5 (14%)
3	70 (9%)	7	45 (64%)	9 (13%)
4	66 (8%)	8	37 (56%)	10 (15%)
5	12 (2%)	2	6 (50%)	3 (25%)
6	16 (2%)	2	7 (44%)	2 (13%)
7	43 (5%)	5	24 (56%)	5 (12%)
8	33 (4%)	5	16 (48%)	6 (18%)
9	17 (2%)	2	10 (59%)	2 (12%)
10	20 (3%)	2	13 (65%)	3 (15%)
11	15 (2%)	1	8 (53%)	3 (20%)
12	33 (4%)	5	17 (52%)	6 (18%)
13	17 (2%)	2	11 (65%)	3 (18%)
14	51 (6%)	6	24 (47%)	12 (24%)
15	25 (3%)	3	16 (64%)	3 (12%)
16	44 (6%)	5	24 (55%)	8 (18%)
17	19 (2%)	2	12 (63%)	2 (11%)
18	16 (2%)	2	8 (50%)	2 (13%)
19	22 (3%)	3	13 (59%)	3 (14%)
20	27 (3%)	2	16 (59%)	2 (7%)
21	28 (4%)	1	13 (46%)	2 (7%)
22	16 (2%)	2	7 (44%)	4 (25%)
23	7 (1%)	1	3 (43%)	1 (14%)
24	93 (12%)	11	48 (52%)	16 (17%)
Total	790	89	435 (55%)	118 (15%)

[1] Percentages refer to the total number of individuals.

[2] Percentages refer to the number of individuals within the family.

* See Figure 13 for example of probands.

In **Table 3**, further details regarding MS cases were reported, including MS course, sex, and age/year of MS onset. The relapse–remitting course (RRMS) was the most represented (49%).

Table 3. Descriptive statistics for the 118 Multiple Sclerosis cases in the Sardinian families.

MS Course °	N (%)	Females (%)	Age MS Onset Mean (SD)	Year MS Onset Mean (SD)
RRMS	58 (49%)	41 (71%)	28.45 (9.49)	1990 (10.09)
SPMS	27 (23%)	14 (52%)	28.89 (8.87)	1983 (9.64)
PPMS	1 (1%)	1 (100%)	45.00	1995
Unknown	32 (27%)	20 (63%)	N/A	N/A
Total	118	76 (64%)	28.64 (9.06) *	1988 (10.88) *

° RRMS = relapse–remitting MS, SPMS = secondary-progressive MS, PPMS = primary-progressive MS, N/A = not available. * A total of 24 subjects had a missing age of MS onset.

In **Table 4**, kinship relationships between the MS-related cases within the families were reported; among all these 238 kinship relationships, the distant relationships over the fourth degree were the most represented, i.e., 176 times (74%), while the other kinship relationships (from the first to the fourth) were found in similar proportions.

Table 4. Kinship relationships between the 118 multiple sclerosis cases.

Kinship Relationship	N (%) *
First degree	20 (8%)
Parent–offspring	9
Mother	6
Father	3
Sibling	13
Second degree	9 (4%)
Uncle/aunt–nephew/niece	8
Grandparent–grandchild	1
Third degree	16 (7%)
Cousins	15
Grand-grandparent–grand-grandchild	1
Fourth degree	17 (7%)
Over the fourth degree	176 (74%)
Total	238

* Percentages refer to the total number of kinship relationships.

3.2.2 Bayesian-LTMH results

3.2.2.1 Primary analysis

The Bayesian-LTMH was implemented including sex and categorized year of birth as covariates, considering all 790 individuals in the Sardinian families. No diagnostic problems were encountered: Gelman-Rubin Statistic (R-hat) was always equal to 1.00 for all parameters, effective sample size was greater than 1000 for each parameter, and traceplots show the convergence of the four chains to the same posterior distribution for each parameter (see **Figure 15**).

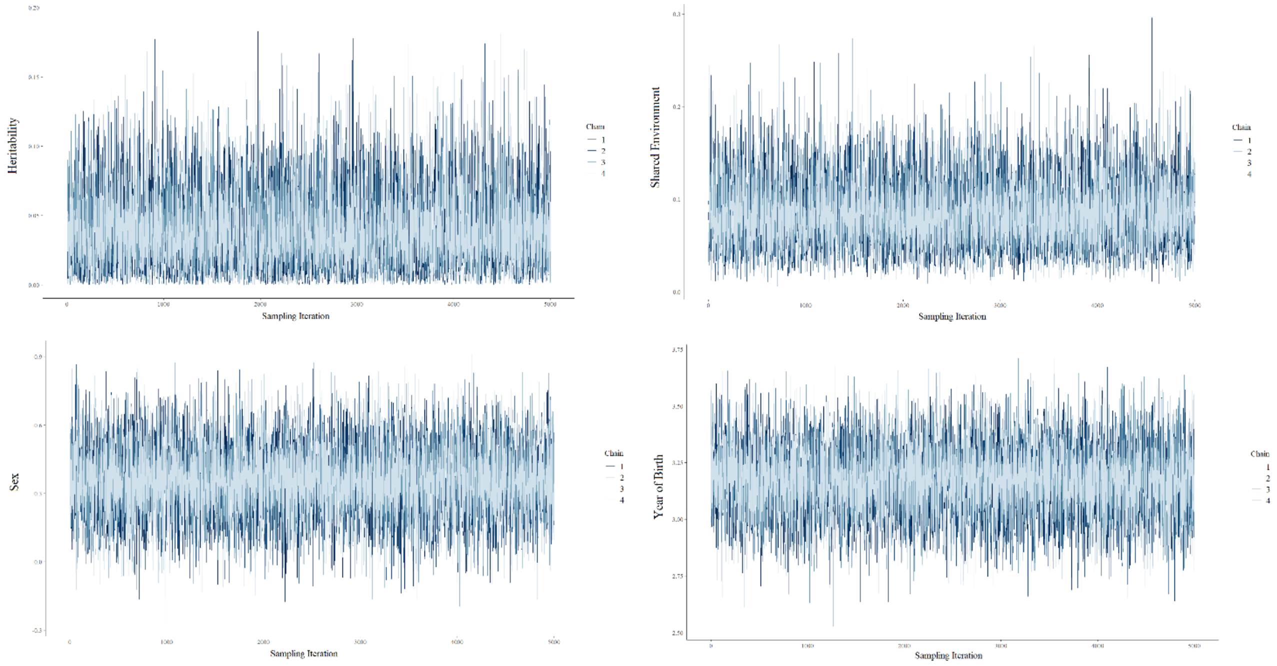


Figure 15. Traceplots showing sampling iterations from the four chains for h^2, c^2_{Total}, β_{SEX} and β_{YR} parameters.

Table 5 reports the results from the first analysis on the whole sample, including the median posterior distributions of the parameters, their standard deviation (SD), and the 95% highest posterior density credibility intervals (HPD CIs). The posterior distributions for the parameters are graphically displayed in **Figure 16**, along with median value (in red) and 95% HPD CIs (in blue).

Table 5. Posterior distributions summary statistics for parameters included in the Bayesian-LTMH applied to the Sardinian families.

Parameter	Median	SD [1]	HPD 95% CI [1]
h^2	0.033	0.028	0.000, 0.094
c^2_{Sibs}	0.033	0.016	0.007, 0.067
$c^2_{Father-Sibs}$	0.012	0.012	0.000, 0.039
$c^2_{Mother-Sibs}$	0.013	0.013	0.000, 0.040
$c^2_{Spouses}$	0.014	0.017	0.000, 0.051
c^2_{Total}	0.080	0.037	0.021, 0.158
e^2	0.168	0.036	0.094, 0.233
$\tau^2_{\beta SEX,YR}$	0.712	0.020	0.673, 0.749
$\tau^2_{\beta SEX}$	0.009	0.008	0.000, 0.027
$\tau^2_{\beta YR}$	0.686	0.024	0.637, 0.731
$2cov°_{\beta SEX,YR}$	0.015	0.007	0.003, 0.028
$\beta_{SEX(Females\ vs.\ Males)}$	0.355	0.157	0.057, 0.679
$\beta_{YR(\geq 1946\ vs.\ <1946)}$	3.173	0.155	2.869, 3.477

[1] SD = standard deviation, HPD 95% CI = highest posterior density 95% credibility interval. Proportion of MS liability variability explained by (i) h^2 = additive genetic effects, (ii) c^2_{Sibs} = siblings' shared environment effects, (iii) $c^2_{Father-Sibs}$ = shared environment effects between the father and the offspring, (iv) $c^2_{Mother-Sibs}$ = shared environment effects between mother and the offspring, (v) $c^2_{Spouses}$ = shared environment effects between spouses, (vi) c^2_{Total} = total shared environment effects, (vii) e^2 = individual environmental effects, (viii) $\tau^2_{\beta SEX,YR}$ = sex and year of birth, (ix) $\tau^2_{\beta SEX}$ = sex, (x) $\tau^2_{\beta YR}$ = year of birth, and (xi) $2cov°_{\beta SEX,YR}$ = covariance between sex and year of birth. β_{SEX} = increase in liability for females compared to males; β_{YR} = increase in liability year of birth on/after 1946 compared to before 1946.

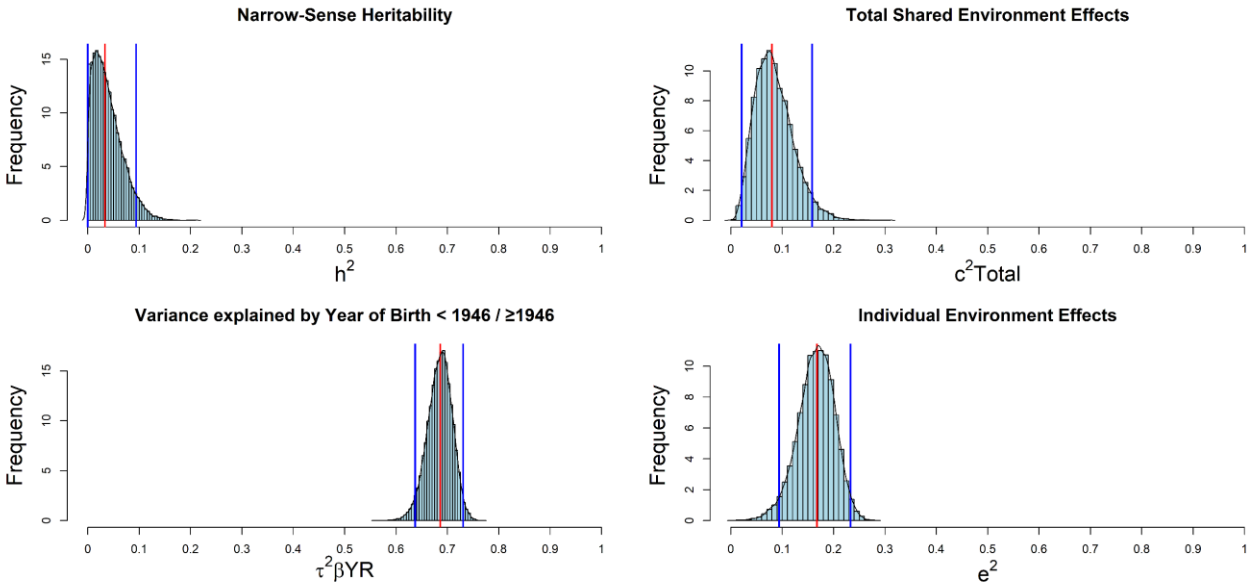


Figure 16. Posterior distributions for parameters included in the Bayesian-LTMH applied to the Sardinian families.

Categorized year of birth resulted as the strongest explanatory factor for MS liability variability, i.e., $\tau^2_{\beta YR} = 0.69$ [95% CI: 0.64, 0.73], meaning that being born before or on/after 1946 explained ~70% of MS liability variability in the Sardinian population. Moreover, compared to individuals born before 1946, individuals born on/after 1946 resulted in a high MS liability increase, i.e., β_{YR} (reference group ≤ 1946) = 3.17 [95% CI: 2.87, 3.48]. This result highlighted year of birth as the major contributor for MS liability variability at the population level, suggesting a crucial role for early environmental exposures. These could be related to the so-called "westernization process", among which different pollution levels, sanitary conditions, and dietary habits other than the malaria eradication program using the insecticide DDT (1946-1950). Early environmental exposures are important to the development and programming of the immune system. Therefore, early immune system interactions can have long-lasting effects on immune responses later in life and potentially contribute to autoimmune diseases like MS. This could be associated with the sudden lack of *Plasmodium falciparum* immune trigger in the environment. Notably, the latter has been hypothesized to be associated with the increasing Sardinian MS incidence and prevalence observed in the last 50 years, following the logic of the hygiene hypothesis [84,88,92,145]. According to this hypothesis, cells of the innate immune system, selected over the centuries to contrast *Plasmodium falciparum* malaria, have kept the tendency to produce abnormal immune responses to new environmental factors triggers, linked to the "westernization process", even after the disappearance of malaria, consequently leading to an increased autoimmune risk.

Individual and shared environmental factors, not linked to the year of birth, explained ~17% ($e^2 = 0.17$ [95% CI: 0.09, 0.23]) and ~8% ($c^2_{Total} = 0.08$ [95% CI: 0.02, 0.16]) of MS liability variability, respectively. These could depend on MS risk factors shared between individuals in the same household or specific to the individual, such as past viral infections (e.g., EBV), smoking habits, exposures to pollutants, low vitamin D levels, dietary habits, and childhood/adolescence obesity [41,146–150].

Genetic variability resulted as a poor explanatory factor, i.e., $h^2 = 0.03$ [95% CI: 0.00, 0.09]. This result implies that genetic variability's contribution in explaining MS liability variability in this specific population is extremely low compared to the other environmental factors. Lastly, sex resulted in a statistically significant increase in MS liability for the "females vs. males" comparison, i.e., β_{SEX} = 0.36 [95% CI: 0.06, 0.68]; however, its explanatory role for MS liability variability was very low compared to the other parameters, i.e., $\tau^2_{\beta SEX}$ median value = 0.01 [95% CI: 0.00, 0.03].

In a separate model, G×E effects variance, i.e., $h^2_{G\times E}$, due to interaction between additive genetics effects and categorized year of birth was also included. The estimated $h^2_{G\times E}$ resulted equal to 0.03 [95% CI: 0.00, 0.10], while categorized year of birth remained the main explanatory factor, i.e., $\tau^2_{\beta YR}$ = 0.69 [95% CI: 0.64, 0.73]. This result implies that the GxE interaction between early environmental exposures and genetic variants has a low explanatory role for MS susceptibility variability at a population level.

Therefore, additive genetic effects, as well as their interaction with the environments linked to the categorized year of birth, did not result as explanatory factors for MS variability. It's important to understand that this result does not imply that genetic variability does not have a causal effect on MS, nor that these additive genetic effects did not significantly change between the two categorized year of birth groups (as highlighted in the following analysis), nor that genetics, in a broader sense, is not involved in determining the disease. Rather, the results imply that genetic variability contribution in explaining MS variability, compared to the other environmental factors, was extremely low [66]. A straightforward consequence is that to have the greatest benefit in preventing MS cases at the population level one should primarily "intervene" on all the factors linked to the year of birth, while "intervening" on the genetic variability would have an extremely lower impact.

From these results, as for all heritability studies, it is not possible to establish causal pathways for the disease onset but, entering in the realm of speculations, it is still possible to provide interesting suggestions for future research. As previously mentioned, the factors linked to the year of birth < or ≥ 1946 are unfortunately unknown based on this analysis but could be potentially linked to the "westernization process". Since these environmental factors represent the primary source for MS variability, the result may provide support for the previously described hygiene hypothesis. Moreover, given the poor explanatory role for GxE effects, there is the indication that the environmental moderation on additive genetic effects had a lower impact on MS variability at population level compared to the direct effects of year of birth (which could have implied other biological pathways).

In summary, we can propose the following hypothesis to elucidate why early environmental exposures play a substantial explanatory role:

1. Malaria is caused by *Plasmodium* parasites and is known to stimulate the immune system [151,152]. Exposure to these parasites in Sardinian individuals had led, over the centuries, to immune-genetic selection and development of immune responses that may have influenced the maturation and regulation of the immune system. The immune system requires balanced exposure to various antigens and challenges during development to establish immune tolerance and prevent autoimmune responses.
2. The hygiene hypothesis suggests that exposures to infections and microbes in early life helps educate the immune system and promote immune tolerance [153]. For example, infections encountered during childhood can "train" the immune system and influence its responsiveness.
3. With the eradication of malaria, a once-prevalent infectious disease that had coexisted with the Sardinian population for centuries, subsequent generations have been deprived of these

immune challenges. The impact of malaria eradication would have then affected those born after the campaign as well as the subsequent generations.

4. In the absence of certain infections or immune challenges, there is a theoretical risk that the immune system may become more prone to autoimmune reactions. These reactions could be triggered by environmental exposures such as viral infections, diet, air quality, pollution, and exposure to allergens. The consequent abnormal immune responses may have led to an increased risk of autoimmune conditions such as MS [92].

While these considerations suggest that malaria eradication could have had implications for immune system development and autoimmune disease risk, it's essential to emphasize that this is a complex area of study, and causative links are challenging to establish definitively. Research into the long-term immunological consequences of the Sardinian environmental changes may provide further insights into the increasing MS risk in this population.

3.2.2.2 Secondary analysis

A secondary analysis was conducted stratifying the sample based on the categorized year of birth, thus focusing on individuals with more similar early environmental exposures. Therefore, the main explanatory role year of birth has been ruled out.

The first group, i.e., "<1946", was composed of 488 subjects: 238 males (49%) and 250 females (51%); 16 MS cases (3%) and 472 healthy controls (97%). The second group, i.e., "≥1946", was instead composed of 302 subjects: 117 males (39%) and 185 females (61%); 102 MS cases (34%) and 200 healthy controls (66%). **Table 6** reports the results from the Bayesian-LTMH model on both groups. The marginal posterior distributions for the parameters are graphically displayed in **Figure 17** for both groups, along with median values (in red) and 95% HPD CIs (in blue).

Table 6. Posterior distributions summary statistics for parameters included in the Bayesian-LTMH applied to the Sardinian families stratified by year of birth on different environment conditions.

	Year of Birth < 1946			**Year of Birth ≥ 1946**		
Parameter	**Median**	**SD** [1]	**95% HPD CI** [1]	**Median**	**SD** [1]	**95% HPD CI** [1]
h^2	0.090	0.100	0.000, 0.312	0.818	0.068	0.679, 0.937
c^2_{Sibs}	0.223	0.100	0.055, 0.433	0.045	0.030	0.004, 0.109
$c^2_{\text{Father–Sibs}}$	0.061	0.058	0.000, 0.185	0.013	0.016	0.000, 0.050
$c^2_{\text{Mother–Sibs}}$	0.049	0.051	0.000, 0.163	0.014	0.017	0.000, 0.054
c^2_{Spouses}	0.085	0.083	0.000, 0.297	0.019	0.026	0.000, 0.078
c^2_{Total}	0.477	0.142	0.199, 0.750	0.105	0.056	0.019, 0.222
e^2	0.086	0.083	0.000, 0.265	0.021	0.025	0.000, 0.078
$\tau^2_{\beta\text{SEX,YR}}$	N/A [1]	N/A [1]	N/A [1]	0.042	0.032	0.000, 0.109
$\tau^2_{\beta\text{SEX}}$	0.304	0.112	0.079, 0.506	0.005	0.013	0.000, 0.035
$\tau^2_{\beta\text{YR}}$	N/A [1]	N/A [1]	N/A [1]	0.032	0.030	0.001, 0.095
$2\text{cov}^\circ_{\beta\text{SEX,YR}}$	N/A [1]	N/A [1]	N/A [1]	0.000	0.001	−0.001, 0.001
$\beta_{\text{SEX(Females vs. Males)}}$	1.322	0.368	0.586, 2.023	0.104	0.177	−0.246, 0.448
$\beta_{\text{YR(10 years increase)}}$	N/A [1]	N/A [1]	N/A [1]	0.186	0.089	0.012, 0.362

[1] SD = standard deviation, HPD = highest posterior density credibility interval, N/A = not available. Proportion of MS liability variability explained by (i) h^2 = additive genetic effects, (ii) c^2_{Sibs} = siblings' shared environment effects, (iii) $c^2_{\text{Father–Sibs}}$ = shared environment effects between the father and the offspring, (iv) $c^2_{\text{Mother–Sibs}}$ = shared environment effects between mother and the offspring, (v) c^2_{Spouses} = shared environment effects between spouses, (vi) c^2_{Total} = total shared environment effects, (vii) e^2 = individual environmental effects, (viii) $\tau^2_{\beta\text{SEX,YR}}$ = sex and year of birth, (ix) $\tau^2_{\beta\text{SEX}}$ = sex, (x) $\tau^2_{\beta\text{YR}}$ = year of birth, and (xi) $2\text{cov}^\circ_{\beta\text{SEX,YR}}$ = covariance between sex and year of birth. β_{SEX} = increase in liability for females compared to males; β_{YR} = increase in liability for 10 years increase in year of birth.

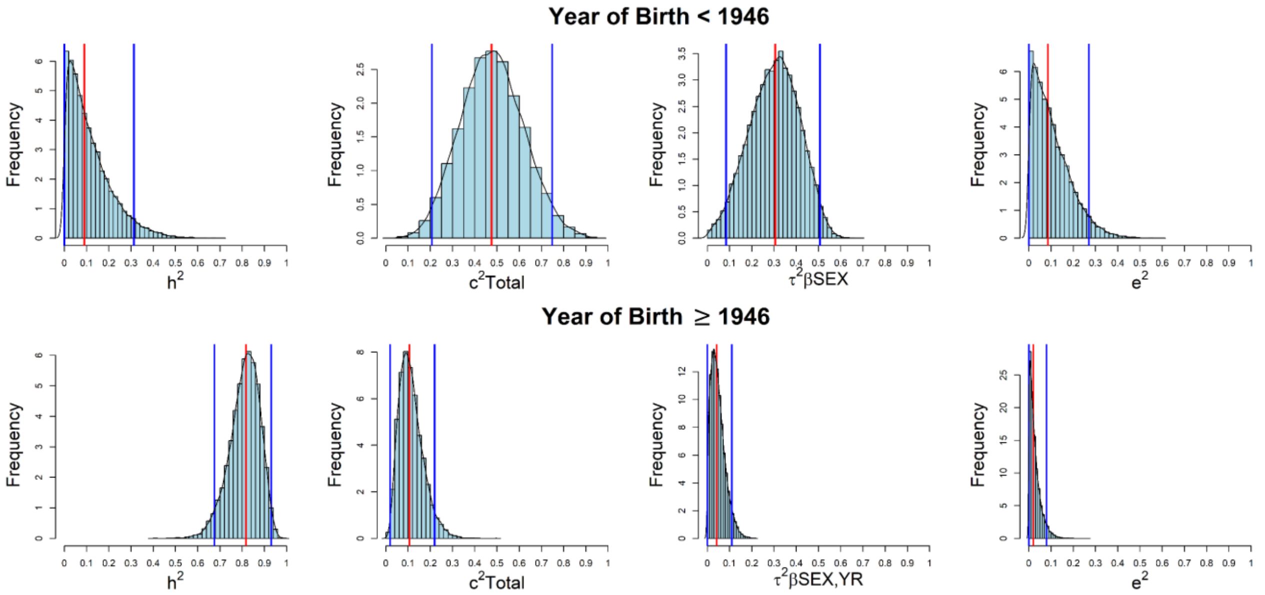


Figure 17. Posterior distributions for parameters included in the Bayesian-LTMH applied to the Sardinian families stratified by year of birth.

The h^2 posterior distribution greatly differed between the two groups, i.e., 0.09 [95% CI: 0.00, 0.31] for the "<1946" group and 0.82 [95% CI: 0.68, 0.93] for the "≥1946" group, indicating that genetic variability acquired a high explanatory role for MS liability variability only considering individuals born on/after 1946. It is worth mentioning that this result does not implicate that MS onset, for an individual born on/after 1946, is caused at 80% by its genetic component, but that observed variability in MS susceptibility in this group is mostly explained by the cumulative effect of an unknown number of risk alleles. For an MS-affected individual born on/after 1946, the high h^2 value provides a strong likelihood that the genetic variability made a greater contribution compared to environmental factors (specific to "year of birth ≥1946" group) in producing a deviation from the population MS liability mean [72]. Potential hypotheses to explain the higher value of h^2 in the second group compared to the first, i.e., (~82% vs. ~9%), could be the following: (i) a decrease in the influence of environmental factors, implying that the genetic variability acquired a higher explanatory role only because the relative explanatory importance was reversed; (ii) an increase in additive genetic effects, implying that the change in environmental factors caused genetic variants to operate differently; (iii) both cases together. The former hypothesis could imply that the effect of certain environmental factors was diminished or were even removed for individuals born after 1946; among these factors, a different medical attention and diagnostic accuracy could provide a potential explanation. Instead, the latter hypothesis may be again aligned with the hygiene hypothesis, as the consequences of malaria eradication could have led genetic variability to gain a higher explanatory role due to its influence on pathways, such as gene expression, related to immune responses and consequently on MS risk [92]. A potential biological mechanism by which early environmental changes influenced additive genetic effects could be found in epigenetics modifications, as these modifications have been shown to have a key role in regulating the expression of immune system genes and have also correlated to autoimmune disorders such as MS [154–156].

Shared environmental effects and sex resulted as the main explanatory components for the "<1946" group, i.e., $c^2_{\text{Total}} = 0.48$ [95% CI: 0.21, 0.75] (with the greatest contribution coming from effects due to environmental factors shared by siblings, c^2_{Sibs} median value = 0.22 [95 CI% = 0.06, 0.43]), and $\tau^2_{\beta\text{SEX}} = 0.31$ [95% CI: 0.08, 0.51]; sex resulted in a statistically significant increase in MS liability for "females vs. males" comparison, i.e., $\beta_{\text{SEX}} = 1.33$ [95% CI: 0.61, 2.03]. Therefore, in this group, specific shared environmental factors (as suggested above), as well as being female, were linked to a higher MS expression at the population level compared to the genetic variability.

Lastly, for the "≥1946" group, it was also possible to include the exact year of birth as a covariate, finding a significantly increasing trend in MS liability, i.e., 0.19 [95% CI: 0.01, 0.36] for an increase of 10 years; however, year of birth explained only ~3% of MS liability variability, i.e., $\tau^2_{\beta\text{YR}} = 0.03$ [95% CI: 0.00, 0.10].

In conclusion, the explanatory sources of MS variability largely differed within the two groups, i.e., shared environmental factors in the "year of birth <1946" group and genetic variability in the "year of birth ≥1946" group, given their different early environmental background.

4 CONCLUSIONS

This work successfully extended the LTMH method, described by Kim, Kwak and Won [96], using a Bayesian framework to estimate MS narrow-sense heritability, on liability scale, making use of Sardinian ascertained family-based samples. LTMH allows adjustment of parameters estimates from ascertainment bias, which occurs when family members are included in the study as relatives of sampled probands. Using Bayesian statistics and MCMC methods, instead of the frequentist EM algorithm-based approach, dramatically improved the computational efficiency of the method, allowing to analyze extended families and improving model flexibility, e.g., including a variance component for shared environment effects. Moreover, considering the lack of a precision measure for the parameters estimates when using EM algorithm, the Bayesian approach allowed to obtain parameters posterior distributions, providing informative and comprehensive results to conduct the statistical inference. To assess the accuracy and the precision of Bayesian-LTMH model, simulation studies were conducted evaluating the obtained posterior distributions for the parameters of interest in different scenarios, comprising pedigree structure, trait's prevalence, and model specifications. Even with non-informative prior distributions, the Bayesian-LTMH provided accurate and reasonably precise posterior distributions, with performance slightly worsening, in terms of bias and precision, for decreasing trait's prevalence.

We then applied Bayesian-LTMH model on a sample of 24 extended pedigrees, ascertained from 89 MS affected probands from the Nuoro province in the Sardinia region. We showed the convergence of the chains to the same posterior distribution and did not encounter any diagnostic problems. We obtained posterior distributions for the proportion of MS susceptibility variability explained by additive genetic effects, shared environment effects, individual environment effects, sex, and year of birth (before or on/after 1946). In line with the latest literature [157], the results pinpoint environmental factors linked to having been born before or on/after 1946 as the leading factors in explaining ~70% of MS liability variability across the 20th century in the Sardinian population. The remaining variability in MS liability (~30%) resulted mainly explained by environmental factors shared among individuals in the same household or specific to the individual (e.g., low vitamin D levels, obesity, past EBV virus infection, diet, and exposure to pollutants). An almost null explanatory role was found for additive genetic effects (i.e., narrow-sense heritability) and GxE effects, suggesting that the genetic variability has low relevance in explaining the sudden increasing MS incidence in the past decades.

Therefore, further investigations would be crucial to identify the specific early environmental factors involved in the increased MS liability in the Sardinian population, as their explanatory role outweighs the role of the other factors. These factors could be researched in the so-called "Westernization process" that took place after World War II, such as different pollution levels, lifestyle, healthcare, and socioeconomic conditions. Another significant and abrupt environmental change occurred with the eradication of malaria in Sardinia between 1946 and 1950 [92]. The Sardinian population has unquestionably experienced selective pressures to develop resistance to malaria, despite carrying a substantial genetic burden. For instance, conditions like thalassemia and glucose-6-phosphate dehydrogenase deficiency have been advantageous genetic adaptations providing protection against *Plasmodium falciparum*, and these genetic variations have become more prevalent in the island due to their effectiveness against the malaria parasite. Interestingly, genetic traits that were favored for their ability to confer protection against *Plasmodium falciparum* infection were also found associated with alleles that predispose individuals to the development of MS. This suggests that in the absence of the influential competitive immune trigger like *Plasmodium*, the generation of individuals born after malaria eradication may have experienced unusually intensified and self-directed responses. These autoimmune responses could have contributed to the notable increase in the incidence of MS observed in Sardinia over the past decades, explaining the main explanatory role of year of birth (before or on/after 1946) in MS susceptibility variability.

Despite the almost null narrow-sense heritability obtained analyzing the whole sample, genetic variability remains a highly relevant matter. In fact, when performing the stratified analysis based on year of birth, genetic variability acquired the main explanatory role for MS liability variability (~82%) in the individuals born on/after 1946. This finding suggests that changes in early environmental exposures after 1946 have led to an increased impact of genetic variability on MS at the population level. This could also be linked to the deleterious effect of genetic variants selected to contrast *Plasmodium* after the disappearance of malaria. Moreover, as highlighted in different studies conducted on Sardinian individuals, genetic variants risk's role has been linked to gene and protein expression, as well as to different peptides structures, which could be involved in causal pathways for MS onset [57,145,158–161]. Therefore, further studies on the Sardinian genetic background could highlight causal biological pathways useful for MS prevention in the current population and for a better understanding of MS etiology.

Comparing heritability estimates between populations, in Sardinian individuals born on or after 1946, it resulted higher (~80%) compared to that obtained using twins from mainland Italy (~50%), Canada (~55%), and the United States (~40%), as well as Finland and France (~25%), while it resulted more similar to h^2 estimates obtained using twins from the United Kingdom (~75%), as well as Denmark and Sweden (~65%) [76]. These results imply that the genetic variability in the Sardinian population, born on or after 1946, has a better explanatory role for MS liability compared to other populations. This could be due to greater additive genetic effects (e.g., specific genetic variants have a higher risk in the Sardinian environmental background), lower environmental effects (e.g., some of the environmental risk factors present in other population may not be part of the Sardinian environmental background), or both. It would be also of interest to integrate year of birth when investigating MS narrow-sense heritability in these different populations, to verify if similar explanatory roles were found for year of birth and genetic variability in absence of a malaria eradication program.

It is worth mentioning that this analysis suffered from some limitations. Firstly, available data did not include other potential confounders, such as smoking habits and previous EBV infection, even if their effect could have been partially captured in the shared environmental effects. Because of the previous limitation, it was not possible to test the explanatory role of other GxE effects which could have confounded heritability estimates. Moreover, the assumed MVN distribution for the underlying liabilities could not be easily checked and, if not respected, could lead to biased estimates [129], therefore further developments are needed to assess the robustness of the results when thi assumption is violated. Nevertheless, the developed Bayesian-LTMH allowed a great advantage to obtain a reasonably precise posterior distribution for MS narrow-sense heritability in the Sardinian population using extended families ascertained from a proband, a result which would have not been achievable with a twin design data due to lack of an adequate sample size [77]. A further strength of this work was represented by the possibility provided by the characteristics of the population under study, i.e., a founder homogenous population, the presence of important environmental risk factors affecting the population such as endemic malaria and its eradication, which allowed to widely investigate the effect of genetic variability, environmental factors, and their interaction MS variability. In conclusion, by discussing pros and cons of heritability studies usefulness and their correct interpretation, a framework was provided to investigate the explanatory role of genetic and environmental factors for other low-prevalence complex traits in specific populations of interest.

Ethics approval and consent to participate

The study was approved by the ethics committee of the Azienda Sanitaria of Nuoro and was conducted in conformity with the 1954 Declaration of Helsinki in its currently applicable version and applicable Italian laws. All study participants gave written informed consent.

Bibliography

1. Fazia, T.; Giulia, |; Baldrighi, N.; Nova, A.; Bernardinelli, L.; Alvarez, F. A systematic review of Mendelian randomization studies on multiple sclerosis. *Eur. J. Neurosci.* **2023**, doi:10.1111/EJN.16088.

2. Nylander, A.; Hafler, D.A. Multiple sclerosis. *J. Clin. Invest.* **2012**, *122*, 1180–1188, doi:10.1172/JCI58649.

3. Goris, A.; Vandebergh, M.; McCauley, J.L.; Saarela, J.; Cotsapas, C. Genetics of multiple sclerosis: lessons from polygenicity. *Lancet Neurol.* **2022**, *21*, 830–842, doi:10.1016/S1474-4422(22)00255-1.

4. Kenealy, S.J.; Pericak-Vance, M.A.; Haines, J.L. The genetic epidemiology of multiple sclerosis. *J. Neuroimmunol.* **2003**, *143*, 7–12, doi:10.1016/j.jneuroim.2003.08.005.

5. Simoens, S. Societal economic burden of multiple sclerosis and cost-effectiveness of disease-modifying therapies. *Front. Neurol.* **2022**, *13*, doi:10.3389/FNEUR.2022.1015256.

6. Kurtzke, J.F. An evaluation of the geographic distribution of multiple sclerosis. *Acta Neurol. Scand.* **1966**, *42*, 91–117, doi:10.1111/J.1600-0404.1966.TB02008.X.

7. Sabel, C.E.; Pearson, J.F.; Mason, D.F.; Willoughby, E.; Abernethy, D.A.; Taylor, B. V. The latitude gradient for multiple sclerosis prevalence is established in the early life course. *Brain* **2021**, *144*, doi:10.1093/brain/awab104.

8. O'Gorman, C.; Lucas, R.; Taylor, B. Environmental risk factors for multiple sclerosis: a review with a focus on molecular mechanisms. *Int. J. Mol. Sci.* **2012**, *13*, 11718–11752, doi:10.3390/IJMS130911718.

9. Urru, S.A.M.; Antonelli, A.; Sechi, G.M. Prevalence of multiple sclerosis in Sardinia: A systematic cross-sectional multi-source survey. *Mult. Scler. J.* **2020**, *26*, 372–380, doi:10.1177/1352458519828600.

10. Montomoli, C.; Prokopenko, I.; Caria, A.; Ferrai, R.; Mander, A.; Seaman, S.; Musu, L.; Piras, M.L.; Ticca, A.F.; Murgia, S.B.; et al. Multiple sclerosis recurrence risk for siblings in an isolated population of Central Sardinia, Italy. *Genet. Epidemiol.* **2002**, *22*, 265–271, doi:10.1002/GEPI.0173.

11. Manjunatha, R.T.; Habib, S.; Sangaraju, S.L.; Yepez, D.; Grandes, X.A. Multiple Sclerosis: Therapeutic Strategies on the Horizon. *Cureus* **2022**, *14*, doi:10.7759/CUREUS.24895.

12. Ford, H. Clinical presentation and diagnosis of multiple sclerosis. *Clin. Med.* **2020**, *20*, 380–383, doi:10.7861/CLINMED.2020-0292.

13. Klineova, S.; Lublin, F.D. Clinical Course of Multiple Sclerosis. *Cold Spring Harb. Perspect. Med.* **2018**, *8*, doi:10.1101/CSHPERSPECT.A028928.

14. Goldman, M.D.; Cross, A.; Riley, C. Treatment of Multiple Sclerosis. *Continuum (Minneap. Minn).* **2022**, *28*, 1025–1051, doi:10.1212/CON.0000000000001170.

15. Ponath, G.; Park, C.; Pitt, D. The role of astrocytes in multiple sclerosis. *Front. Immunol.* **2018**, *9*, 337102, doi:10.3389/FIMMU.2018.00217/BIBTEX.

16. Dighriri, I.M.; Aldalbahi, A.A.; Albeladi, F.; Tahiri, A.A.; Kinani, E.M.; Almohsen, R.A.; Alamoudi, N.H.; Alanazi, A.A.; Alkhamshi, S.J.; Althomali, N.A.; et al. An Overview of the History, Pathophysiology, and Pharmacological Interventions of Multiple Sclerosis. *Cureus*

2023, *15*, doi:10.7759/CUREUS.33242.

17. Ghasemi, N.; Razavi, S.; Nikzad, E. Multiple Sclerosis: Pathogenesis, Symptoms, Diagnoses andCell-Based Therapy. *Cell J.* **2017**, *19*, 1, doi:10.22074/CELLJ.2016.4867.

18. Nourbakhsh, B.; Mowry, E.M. Multiple sclerosis risk factors and pathogenesis. *Contin. Lifelong Learn. Neurol.* 2019, *25*, doi:10.1212/CON.0000000000000725.

19. Alvarez, J.I.; Cayrol, R.; Prat, A. Disruption of central nervous system barriers in multiple sclerosis. *Biochim. Biophys. Acta - Mol. Basis Dis.* 2011, *1812*, doi:10.1016/j.bbadis.2010.06.017.

20. Petrova, N.; Carassiti, D.; Altmann, D.R.; Baker, D.; Schmierer, K. Axonal loss in the multiple sclerosis spinal cord revisited. *Brain Pathol.* **2018**, *28*, doi:10.1111/bpa.12516.

21. Chastain, E.M.L.; Duncan, D.S.; Rodgers, J.M.; Miller, S.D. The role of antigen presenting cells in multiple sclerosis. *Biochim. Biophys. Acta - Mol. Basis Dis.* 2011, *1812*, doi:10.1016/j.bbadis.2010.07.008.

22. Miterski, B.; Böhringer, S.; Klein, W.; Sindern, E.; Haupts, M.; Schimrigk, S.; Epplen, J.T. Inhibitors in the NFkappaB cascade comprise prime candidate genes predisposing to multiple sclerosis, especially in selected combinations. *Genes Immun.* **2002**, *3*, 211–219, doi:10.1038/SJ.GENE.6363846.

23. Lassmann, H.; Brück, W.; Lucchinetti, C.F. The immunopathology of multiple sclerosis: an overview. *Brain Pathol.* **2007**, *17*, 210–218, doi:10.1111/J.1750-3639.2007.00064.X.

24. Hillert, J.; Masterman, T. The Genetics of Multiple Sclerosis. *Handb. Mult. Sclerosis, Third Ed.* **2017**, 33–65, doi:10.15586/CODON.MULTIPLESCLEROSIS.2017.CH1.

25. Wheeler, M.A.; Quintana, F.J. Regulation of astrocyte functions in multiple sclerosis. *Cold Spring Harb. Perspect. Med.* **2019**, *9*, doi:10.1101/cshperspect.a029009.

26. Yue, Y.; Stone, S.; Lin, W. Role of nuclear factor κB in multiple sclerosis and experimental autoimmune encephalomyelitis. *Neural Regen. Res.* **2018**, *13*, 1507, doi:10.4103/1673-5374.237109.

27. Chaudhuri, A. Multiple sclerosis is primarily a neurodegenerative disease. *J. Neural Transm.* **2013**, *120*, 1463–1466, doi:10.1007/S00702-013-1080-3.

28. Poser, C.M.; Paty, D.W.; Scheinberg, L.; McDonald, W.I.; Davis, F.A.; Ebers, G.C.; Johnson, K.P.; Sibley, W.A.; Silberberg, D.H.; Tourtellotte, W.W. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann. Neurol.* **1983**, *13*, 227–231, doi:10.1002/ANA.410130302.

29. Nielsen, N.M.; Westergaard, T.; Rostgaard, K.; Frisch, M.; Hjalgrim, H.; Wohlfahrt, J.; Koch-Henriksen, N.; Melbye, M. Familial risk of multiple sclerosis: A nationwide cohort study. *Am. J. Epidemiol.* **2005**, *162*, doi:10.1093/aje/kwi280.

30. Owens, G.P.; Gilden, D.; Burgoon, M.P.; Yu, X.; Bennett, J.L. Viruses and multiple sclerosis. *Neuroscientist* **2011**, *17*, 659–676, doi:10.1177/1073858411386615.

31. Gale, C.R.; Martyn, C.N. Migrant studies in multiple sclerosis. *Prog. Neurobiol.* 1995, *47*, doi:10.1016/0301-0082(95)80008-V.

32. Hayes, C.E.; Cantorna, M.T.; DeLuca, H.F. Vitamin D and multiple sclerosis. *Proc. Soc. Exp. Biol. Med.* **1997**, *216*, 21–27, doi:10.3181/00379727-216-44153A.

33. Sintzel, M.B.; Rametta, M.; Reder, A.T. Vitamin D and Multiple Sclerosis: A Comprehensive Review. *Neurol. Ther.* **2018**, *7*, 59, doi:10.1007/S40120-017-0086-4.

34. Matías-Guíu, J.; Oreja-Guevara, C.; Matias-Guiu, J.A.; Gomez-Pinedo, U. Vitamin D and remyelination in multiple sclerosis. *Neurologia* 2018, *33*, doi:10.1016/j.nrl.2016.05.001.

35. Wang, W.; Li, Y.; Meng, X. Vitamin D and neurodegenerative diseases. *Heliyon* 2023, *9*, doi:10.1016/j.heliyon.2023.e12877.

36. Mokry, L.E.; Ross, S.; Ahmad, O.S.; Forgetta, V.; Smith, G.D.; Leong, A.; Greenwood, C.M.T.; Thanassoulis, G.; Richards, J.B. Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. *PLOS Med.* **2015**, *12*, e1001866, doi:10.1371/JOURNAL.PMED.1001866.

37. González-Madrid, E.; Rangel-Ramírez, M.A.; Mendoza-León, M.J.; Álvarez-Mardones, O.; González, P.A.; Kalergis, A.M.; Opazo, M.C.; Riedel, C.A. Risk Factors from Pregnancy to Adulthood in Multiple Sclerosis Outcome. *Int. J. Mol. Sci.* **2022**, *23*, doi:10.3390/IJMS23137080.

38. Dörr, J.; Ohlraun, S.; Skarabis, H.; Paul, F. Efficacy of vitamin D supplementation in multiple sclerosis (EVIDIMS Trial): study protocol for a randomized controlled trial. *Trials* **2012**, *13*, doi:10.1186/1745-6215-13-15.

39. Feige, J.; Moser, T.; Bieler, L.; Schwenker, K.; Hauer, L.; Sellner, J. Vitamin D Supplementation in Multiple Sclerosis: A Critical Analysis of Potentials and Threats. *Nutrients* **2020**, *12*, doi:10.3390/NU12030783.

40. Bar-Or, A.; Pender, M.P.; Khanna, R.; Steinman, L.; Hartung, H.P.; Maniar, T.; Croze, E.; Aftab, B.T.; Giovannoni, G.; Joshi, M.J. Epstein–Barr Virus in Multiple Sclerosis: Theory and Emerging Immunotherapies. *Trends Mol. Med.* 2020, *26*, doi:10.1016/j.molmed.2019.11.003.

41. Handel, A.E.; Williamson, A.J.; Disanto, G.; Handunnetthi, L.; Giovannoni, G.; Ramagopalan, S. V. An updated meta-analysis of risk of multiple sclerosis following infectious mononucleosis. *PLoS One* **2010**, *5*, 1–5, doi:10.1371/JOURNAL.PONE.0012496.

42. Aloisi, F.; Giovannoni, G.; Salvetti, M. Epstein-Barr virus as a cause of multiple sclerosis: opportunities for prevention and therapy. *Lancet Neurol.* **2023**, *22*, 338–349, doi:10.1016/S1474-4422(22)00471-9.

43. Manouchehrinia, A.; Hedström, A.K.; Alfredsson, L.; Olsson, T.; Hillert, J.; Ramanujam, R. Association of pre-disease body mass index with multiple sclerosis prognosis. *Front. Neurol.* **2018**, *9*, doi:10.3389/fneur.2018.00232.

44. Dardiotis, E.; Tsouris, Z.; Aslanidou, P.; Aloizou, A.M.; Sokratous, M.; Provatas, A.; Siokas, V.; Deretzi, G.; Hadjigeorgiou, G.M. Body mass index in patients with Multiple Sclerosis: a meta-analysis. *Neurol. Res.* 2019, *41*, doi:10.1080/01616412.2019.1622873.

45. Gianfrancesco, M.A.; Glymour, M.M.; Walter, S.; Rhead, B.; Shao, X.; Shen, L.; Quach, H.; Hubbard, A.; Jónsdóttir, I.; Stefánsson, K.; et al. Causal Effect of Genetic Variants Associated with Body Mass Index on Multiple Sclerosis Susceptibility. *Am. J. Epidemiol.* **2017**, *185*, doi:10.1093/aje/kww120.

46. Rosso, M.; Chitnis, T. Association between Cigarette Smoking and Multiple Sclerosis: A Review. *JAMA Neurol.* 2020, *77*, doi:10.1001/jamaneurol.2019.4271.

47. Minagar, A.; Alexander, J.S. Blood-brain barrier disruption in multiple sclerosis. *Mult. Scler.*

2003, *9*, 540–549, doi:10.1191/1352458503MS965OA.

48. Blomgren, K.; Hagberg, H. Free radicals, mitochondria, and hypoxia-ischemia in the developing brain. *Free Radic. Biol. Med.* **2006**, *40*, 388–397, doi:10.1016/J.FREERADBIOMED.2005.08.040.

49. Handel, A.E.; Williamson, A.J.; Disanto, G.; Dobson, R.; Giovannoni, G.; Ramagopalan, S. V. Smoking and multiple sclerosis: an updated meta-analysis. *PLoS One* **2011**, *6*, doi:10.1371/JOURNAL.PONE.0016149.

50. Preiningerova, J.L.; Zakostelska, Z.J.; Srinivasan, A.; Ticha, V.; Kovarova, I.; Kleinova, P.; Tlaskalova-Hogenova, H.; Havrdova, E.K. Multiple Sclerosis and Microbiome. *Biomolecules* 2022, *12*, doi:10.3390/biom12030433.

51. Jangi, S.; Gandhi, R.; Cox, L.M.; Li, N.; Von Glehn, F.; Yan, R.; Patel, B.; Mazzola, M.A.; Liu, S.; Glanz, B.L.; et al. Alterations of the human gut microbiome in multiple sclerosis. *Nat. Commun.* **2016**, *7*, doi:10.1038/ncomms12015.

52. Katz Sand, I. The Role of Diet in Multiple Sclerosis: Mechanistic Connections and Current Evidence. *Curr. Nutr. Rep.* 2018, *7*, doi:10.1007/s13668-018-0236-z.

53. Gourraud, P.A.; Harbo, H.F.; Hauser, S.L.; Baranzini, S.E. The genetics of multiple sclerosis: An up-to-date review. *Immunol. Rev.* **2012**, *248*, 87, doi:10.1111/J.1600-065X.2012.01134.X.

54. Patsopoulos, N.A.; Baranzini, S.E.; Santaniello, A.; Shoostari, P.; Cotsapas, C.; Wong, G.; Beecham, A.H.; James, T.; Replogle, J.; Vlachos, I.S.; et al. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **2019**, *365*, doi:10.1126/SCIENCE.AAV7188.

55. Housley, W.J.; Fernandez, S.D.; Vera, K.; Murikinati, S.R.; Grutzendler, J.; Cuerdon, N.; Glick, L.; De Jager, P.L.; Mitrovic, M.; Cotsapas, C.; et al. Genetic variants associated with autoimmunity drive NFκB signaling and responses to inflammatory stimuli. *Sci. Transl. Med.* **2015**, *7*, doi:10.1126/SCITRANSLMED.AAA9223.

56. Sun, X.F.; Zhang, H. NFKB and NFKBI polymorphisms in relation to susceptibility of tumour and other diseases. *Histol. Histopathol.* 2007, *22*, doi:10.14670/HH-22.1387.

57. Fazia, T.; Nova, A.; Gentilini, D.; Beecham, A.; Piras, M.; Saddi, V.; Ticca, A.; Bitti, P.; McCauley, J.L.; Berzuini, C.; et al. Investigating the Causal Effect of Brain Expression of CCL2, NFKB1, MAPK14, TNFRSF1A, CXCL10 Genes on Multiple Sclerosis: A Two-Sample Mendelian Randomization Approach. *Front. Bioeng. Biotechnol.* **2020**, *8*, doi:10.3389/FBIOE.2020.00397.

58. Yi, H.; Bai, Y.; Zhu, X.; lin, L.; Zhao, L.; Wu, X.; Buch, S.; Wang, L.; Chao, J.; Yao, H. IL-17A Induces MIP-1α Expression in Primary Astrocytes via Src/MAPK/PI3K/NF-kB Pathways: Implications for Multiple Sclerosis. *J. Neuroimmune Pharmacol.* **2014**, *9*, doi:10.1007/s11481-014-9553-1.

59. Kumar, N.; Sharma, N.; Mehan, S. Connection between JAK/STAT and PPARγ Signaling During the Progression of Multiple Sclerosis: Insights into the Modulation of T-Cells and Immune Responses in the Brain. *Curr. Mol. Pharmacol.* **2021**, *14*, doi:10.2174/1874467214666210301121432.

60. Heinrich, P.C.; Behrmann, I.; Haan, S.; Hermanns, H.M.; Müller-Newen, G.; Schaper, F. Principles of interleukin (IL)-6-type cytokine signalling and its regulation. *Biochem. J.* 2003,

374, doi:10.1042/BJ20030407.

61. Damotte, V.; Guillot-Noel, L.; Patsopoulos, N.A.; Madireddy, L.; El Behi, M.; De Jager, P.L.; Baranzini, S.E.; Cournu-Rebeix, I.; Fontaine, B. A gene pathway analysis highlights the role of cellular adhesion molecules in multiple sclerosis susceptibility. *Genes Immun.* **2014**, *15*, 126–132, doi:10.1038/GENE.2013.70.

62. Bird, A. Perceptions of epigenetics. *Nat. 2007 4477143* **2007**, *447*, 396–398, doi:10.1038/nature05913.

63. Miyazaki, Y.; Niino, M. Epigenetics in multiple sclerosis. *Clin. Exp. Neuroimmunol.* 2015, *6*, doi:10.1111/cen3.12271.

64. Chan, V.S.F. Epigenetics in Multiple Sclerosis. In *Advances in Experimental Medicine and Biology*; 2020; Vol. 1253.

65. Van Den Elsen, P.J.; Van Eggermond, M.C.J.A.; Puentes, F.; Van Der Valk, P.; Baker, D.; Amor, S. The epigenetics of multiple sclerosis and other related disorders. *Mult. Scler. Relat. Disord.* 2014, *3*, doi:10.1016/j.msard.2013.08.007.

66. Pearson, C.H. Is heritability explanatorily useful? *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2007**, *38*, 270–288, doi:10.1016/J.SHPSC.2006.12.012.

67. Kempthorne, O. The correlation between relatives on the supposition of mendelian inheritance. *Am. J. Hum. Genet.* **1968**, *20*, 402.

68. Athanasiadis, G.; Speed, D.; Andersen, M.K.; Appel, E.V.R.; Grarup, N.; Brandslund, I.; Jørgensen, M.E.; Larsen, C.V.L.; Bjerregaard, P.; Hansen, T.; et al. Estimating narrow-sense heritability using family data from admixed populations. *Hered. 2020 1246* **2020**, *124*, 751–762, doi:10.1038/s41437-020-0311-2.

69. Visscher, P.M.; Hill, W.G.; Wray, N.R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet. 2008 94* **2008**, *9*, 255–266, doi:10.1038/nrg2322.

70. Uchiyama, R.; Spicer, R.; Muthukrishna, M. Cultural evolution of genetic heritability. *Behav. Brain Sci.* **2022**, *45*, e152, doi:10.1017/S0140525X21000893.

71. Bourrat, P. Heritability, causal influence and locality. *Synthese* **2021**, *198*, 6689–6715, doi:10.1007/S11229-019-02484-3/FIGURES/4.

72. Tal, O. From heritability to probability. *Biol. Philos.* **2009**, *24*, 81–105, doi:10.1007/S10539-008-9129-7/FIGURES/8.

73. Milo, R.; Kahana, E. Multiple sclerosis: geoepidemiology, genetics and the environment. *Autoimmun. Rev.* **2010**, *9*, doi:10.1016/J.AUTREV.2009.11.010.

74. Compston, A.; Coles, A. Multiple sclerosis. *Lancet (London, England)* **2008**, *372*, 1502–1517, doi:10.1016/S0140-6736(08)61620-7.

75. Egeland, J. Heritability and Etiology: Heritability estimates can provide causally relevant information. *Pers. Individ. Dif.* **2023**, *200*, 111896, doi:10.1016/J.PAID.2022.111896.

76. Fagnani, C.; Neale, M.C.; Nisticò, L.; Stazi, M.A.; Ricigliano, V.A.; Buscarinu, M.C.; Salvetti, M.; Ristori, G. Twin studies in multiple sclerosis: A meta-estimation of heritability and environmentality. *Mult. Scler.* **2015**, *21*, 1404–1413, doi:10.1177/1352458514564492.

77. Hawkes, C.H.; Macgregor, A.J. Twin studies and the heritability of MS: a conclusion. *Mult. Scler.* **2009**, *15*, 661–667, doi:10.1177/1352458509104592.

78. ISTAT Estimated resident population - Years 2002-2019 : Sardegna 3 Available online: http://dati.istat.it/Index.aspx?QueryId=12410&lang=en (accessed on Jan 25, 2023).

79. Ristori, G.; Cannoni, S.; Stazi, M.A.; Vanacore, N.; Cotichini, R.; Alfò, M.; Pugliatti, M.; Sotgiu, S.; Solaro, C.; Bomprezzi, R.; et al. Multiple Sclerosis in Twins from Continental Italy and Sardinia: A Nationwide Study. **2005**, doi:10.1002/ana.20683.

80. Docherty, A.R.; Kremen, W.S.; Panizzon, M.S.; Prom-Wormley, E.C.; Franz, C.E.; Lyons, M.J.; Eaves, L.J.; Neale, M.C. Comparison of Twin and Extended Pedigree Designs for Obtaining Heritability Estimates. *Behav. Genet.* **2015**, *45*, 461, doi:10.1007/S10519-015-9720-Z.

81. Kruuk, L.E.B.; Hadfield, J.D. How to separate genetic and environmental causes of similarity between relatives. *J. Evol. Biol.* **2007**, *20*, 1890–1903, doi:10.1111/J.1420-9101.2007.01377.X.

82. Dick, D.M. Shared Environment. *Encycl. Stat. Behav. Sci.* **2005**, doi:10.1002/0470013192.BSA611.

83. Pittner, K.; Bakermans-Kranenburg, M.J.; Alink, L.R.A.; Buisman, R.S.M.; van den Berg, L.J.M.; Block, L.H.C.G.C.C. de; Voorthuis, A.; Elzinga, B.M.; Lindenberg, J.; Tollenaar, M.S.; et al. Estimating the Heritability of Experiencing Child Maltreatment in an Extended Family Design. *Child Maltreat.* **2020**, *25*, 289, doi:10.1177/1077559519888587.

84. Granieri, E.; Casetta, I.; Govoni, V.; Tola, M.R.; Marchi, D.; Murgia, S.B.; Ticca, A.; Pugliatti, M.; Murgia, B.; Rosati, G. The increasing incidence and prevalence of MS in a Sardinian province. *Neurology* **2000**, *55*, 842–848, doi:10.1212/WNL.55.6.842.

85. Sotgiu, S.; Pugliatti, M.; Sotgiu, A.; Sanna, A.; Rosati, G. Review: Does the "Hygiene Hypothesis" Provide an Explanation for the High Prevalence of Multiple Sclerosis in Sardinia? *http://dx.doi.org/10.1080/08916930310001151607* **2009**, *36*, 257–260, doi:10.1080/08916930310001151607.

86. Sotgiu, S.; Pugliatti, M.; Sanna, A.; Sotgiu, A.; Castigli, P.; Solinas, G.; Dolei, A.; Serra, C.; Bonetti, B.; Rosati, G. Multiple sclerosis complexity in selected populations: the challenge of Sardinia, insular Italy. *Eur. J. Neurol.* **2002**, *9*, 329–341, doi:10.1046/J.1468-1331.2002.00412.X.

87. Matveeva, O.; Bogie, J.F.J.; Hendriks, J.J.A.; Linker, R.A.; Haghikia, A.; Kleinewietfeld, M. Western lifestyle and immunopathology of multiple sclerosis. *Ann. N. Y. Acad. Sci.* **2018**, *1417*, 71, doi:10.1111/NYAS.13583.

88. Tognotti, E. Program to Eradicate Malaria in Sardinia, 1946–1950. *Emerg. Infect. Dis.* **2009**, *15*, 1460, doi:10.3201/EID1509.081317.

89. Riedl, B.; Beckmann, T.; Neundõrfer, B.; Handwerker, H.O.; Birklein, F. Multiple sclerosis epidemiology in Sardinia: evidence for a true increasing risk. *Acta Neurol. Scand.* **2001**, *103*, 20–26, doi:10.1034/J.1600-0404.2001.00207.X.

90. Casetta, I.; Granieri, E.; Marchi, D.; Murgia, S.B.; Tola, M.R.; Ticca, A.; Lauria, G.; Govoni, V.; Murgia, B.; Pugliatti, M. An epidemiological study of multiple sclerosis in central Sardinia, Italy. *Acta Neurol. Scand.* **1998**, *98*, 391–394, doi:10.1111/J.1600-0404.1998.TB07319.X.

91. Fleming, J.O.; Cook, T.D. Multiple sclerosis and the hygiene hypothesis. *Neurology* **2006**, *67*, doi:10.1212/01.wnl.0000247663.40297.2d.

92. Sotgiu, S.; Angius, A.; Embry, A.; Rosati, G.; Musumeci, S. Hygiene hypothesis: innate immunity, malaria and multiple sclerosis. *Med. Hypotheses* **2008**, *70*, 819–825, doi:10.1016/J.MEHY.2006.10.069.

93. Pugliatti, M.; Solinas, G.; Sotgiu, S.; Castiglia, P.; Rosati, G. Multiple sclerosis distribution in northern Sardinia: Spatial cluster analysis of prevalence. *Neurology* **2002**, *58*, doi:10.1212/WNL.58.2.277.

94. De Villemereuil, P.; Gimenez, O.; Doligez, B. Comparing parent–offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: a simulation study for Gaussian and binary traits. *Methods Ecol. Evol.* **2013**, *4*, 260–275, doi:10.1111/2041-210X.12011.

95. Park, S.; Lee, S.; Lee, Y.; Herold, C.; Hooli, B.; Mullin, K.; Park, T.; Park, C.; Bertram, L.; Lange, C.; et al. Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families. *BMC Med. Genet.* **2015**, *16*, doi:10.1186/S12881-015-0198-6.

96. Kim, W.; Kwak, S.H.; Won, S. Heritability estimation of dichotomous phenotypes using a liability threshold model on ascertained family-based samples. *Genet. Epidemiol.* **2019**, *43*, 761–775, doi:10.1002/GEPI.22244.

97. Fazia, T.; Pastorino, R.; Foco, L.; Han, L.; Abney, M.; Beecham, A.; Hadjixenofontos, A.; Guo, H.; Gentilini, D.; Papachristou, C.; et al. Investigating multiple sclerosis genetic susceptibility on the founder population of east-central Sardinia via association and linkage analysis of immune-related loci. *Mult. Scler.* **2018**, *24*, 1815–1824, doi:10.1177/1352458517732841.

98. McLachlan, G.J.; Krishnan, T. *The EM Algorithm and Extensions: Second Edition*; 2007;

99. Louis, T.A. Finding the Observed Information Matrix When Using the EM Algorithm. *J. R. Stat. Soc. Ser. B* **1982**, *44*, 226–233, doi:10.1111/J.2517-6161.1982.TB01203.X.

100. Xu, C.; Baines, P.D.; Wang, J.L. Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data. *Biostatistics* **2014**, *15*, 731, doi:10.1093/BIOSTATISTICS/KXU015.

101. Sofer, T. Confidence intervals for heritability via Haseman-Elston regression. *Stat. Appl. Genet. Mol. Biol.* **2017**, *16*, 259, doi:10.1515/SAGMB-2016-0076.

102. Paap, R. What are the advantages of MCMC based inference in latent variable models? *Stat. Neerl.* **2002**, *56*, 2–22, doi:10.1111/1467-9574.00060.

103. van de Schoot, R.; Depaoli, S.; King, R.; Kramer, B.; Märtens, K.; Tadesse, M.G.; Vannucci, M.; Gelman, A.; Veen, D.; Willemsen, J.; et al. Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* 2021, *1*, doi:10.1038/s43586-020-00001-2.

104. Ellison, A.M. Bayesian inference in ecology. *Ecol. Lett.* 2004, *7*, doi:10.1111/j.1461-0248.2004.00603.x.

105. Kruschke, J.K. Bayesian Analysis Reporting Guidelines. *Nat. Hum. Behav.* 2021, *5*, doi:10.1038/s41562-021-01177-7.

106. Hamra, G.; MacLehose, R.; Richardson, D. Markov Chain Monte Carlo: an introduction for epidemiologists. *Int. J. Epidemiol.* **2013**, *42*, 627, doi:10.1093/IJE/DYT043.

107. Chib, S.; Greenberg, E. Understanding the metropolis-hastings algorithm. *Am. Stat.* **1995**, *49*, doi:10.1080/00031305.1995.10476177.

108. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian data analysis, third edition*; 2013;

109. Buchholz, A.; Chopin, N.; Jacob, P.E. Adaptive Tuning of Hamiltonian Monte Carlo Within Sequential Monte Carlo. *Bayesian Anal.* **2021**, *16*, doi:10.1214/20-BA1222.

110. Choudhary, A.; Lindner, J.F.; Holliday, E.G.; Miller, S.T.; Sinha, S.; Ditto, W.L. Forecasting Hamiltonian dynamics without canonical coordinates. *Nonlinear Dyn.* **2021**, *103*, doi:10.1007/s11071-020-06185-2.

111. Betancourt, M. The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo. *Ann. Phys.* **2019**, *531*, doi:10.1002/andp.201700214.

112. Vijayarakavan, M.; Prabhavathi, K. Numerical investigation for the inhomogeneous problems by using leapfrog method. *Adv. Math. Sci. J.* **2020**, *9*, doi:10.37418/amsj.9.5.58.

113. Hoffman, M.D.; Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.

114. Visscher, P.M.; Wray, N.R. Concepts and Misconceptions about the Polygenic Additive Model Applied to Disease. *Hum. Hered.* **2015**, *80*, 165–170, doi:10.1159/000446931.

115. Blangero, J.; Williams, J.T.; Almasy, L. Variance component methods for detecting complex trait loci. *Adv. Genet.* **2001**, *42*, 151–181, doi:10.1016/S0065-2660(01)42021-9.

116. de Villemereuil, P.; Morrissey, M.B.; Nakagawa, S.; Schielzeth, H. Fixed-effect variance and the estimation of repeatabilities and heritabilities: issues and solutions. *J. Evol. Biol.* **2018**, *31*, 621–632, doi:10.1111/JEB.13232.

117. Chi, P.B.; Duncan, A.E.; Kramer, P.A.; Minin, V.N. Heritability estimation of osteoarthritis in the pig-tailed macaque (Macaca Nemestrina) with a look toward future data collection. *PeerJ* **2014**, *2014*, doi:10.7717/PEERJ.373/SUPP-1.

118. Keller, M.C.; Medland, S.E.; Duncan, L.E.; Hatemi, P.K.; Neale, M.C.; Maes, H.H.M.; Eaves, L.J. Modeling extended twin family data I: description of the Cascade model. *Twin Res. Hum. Genet.* **2009**, *12*, 8–18, doi:10.1375/TWIN.12.1.8.

119. Gjessing, H.K.; Lie, R.T. Biometrical modelling in genetics: are complex traits too complex? *Stat. Methods Med. Res.* **2008**, *17*, 75–96, doi:10.1177/0962280207081241.

120. Xia, C.; Amador, C.; Huffman, J.; Trochet, H.; Campbell, A.; Porteous, D.; Hastie, N.D.; Hayward, C.; Vitart, V.; Navarro, P.; et al. Pedigree- and SNP-Associated Genetics and Recent Environment are the Major Contributors to Anthropometric and Cardiometabolic Trait Variation. *PLoS Genet.* **2016**, *12*, doi:10.1371/JOURNAL.PGEN.1005804.

121. Fernández-Rhodes, L.; Howard, A.G.; Tao, R.; Young, K.L.; Graff, M.; Aiello, A.E.; North, K.E.; Justice, A.E. Characterization of the contribution of shared environmental and genetic factors to metabolic syndrome methylation heritability and familial correlations 06 Biological Sciences 0604 Genetics. *BMC Genet.* **2018**, *19*, 7–14, doi:10.1186/S12863-018-0634-7/FIGURES/2.

122. Neale, M.C.; Cardon, L.R. Methodology for Genetic Studies of Twins and Families. *Methodol. Genet. Stud. Twins Fam.* **1992**, doi:10.1007/978-94-015-8018-2.

123. Pilia, G.; Chen, W.M.; Scuteri, A.; Orrú, M.; Albai, G.; Dei, M.; Lai, S.; Usala, G.; Lai, M.; Loi, P.; et al. Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLOS*

Genet. **2006**, *2*, e132, doi:10.1371/JOURNAL.PGEN.0020132.

124. Hill, W.; Arslan, R.C.; Xia, C.; Luciano, M.; Amador, C.; Navarro, P.; Hayward, C.; Nagy, R.; Porteous, D.; Mcintosh, A.M.; et al. Genomic analysis of family data reveals additional genetic effects on intelligence and personality. *Mol. Psychiatry* **2018**, *23*, 2347–2362, doi:10.1038/s41380-017-0005-1.

125. Zhu, Z.; Bakshi, A.; Vinkhuyzen, A.A.E.; Hemani, G.; Lee, S.H.; Nolte, I.M.; Van Vliet-Ostaptchouk, J. V.; Snieder, H.; Esko, T.; Milani, L.; et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am. J. Hum. Genet.* **2015**, *96*, 377–385, doi:10.1016/J.AJHG.2015.01.001.

126. Rabe-Hesketh, S.; Skrondal, A.; Gjessing, H.K. Biometrical Modeling of Twin and Family Data Using Standard Mixed Model Software. *Biometrics* **2008**, *64*, 280–288, doi:10.1111/j.1541-0420.2007.00803.x.

127. Reimherr, M.; Nicolae, D. Estimating Variance Components in Functional Linear Models With Applications to Genetic Heritability. *https://doi.org/10.1080/01621459.2015.1016224* **2016**, *111*, 407–422, doi:10.1080/01621459.2015.1016224.

128. Willemsen, G.; Vink, J.M.; Boomsma, D.I. Assortative mating may explain spouses' risk of same disease. *BMJ* **2003**, *326*, 396, doi:10.1136/BMJ.326.7385.396/A.

129. Benchek, P.H.; Morris, N.J. How meaningful are heritability estimates of liability? *Hum. Genet.* **2013**, *132*, 1351–1360, doi:10.1007/S00439-013-1334-Z.

130. Jacobs, B.M.; Noyce, A.J.; Bestwick, J.; Belete, D.; Giovannoni, G.; Dobson, R. Gene-Environment Interactions in Multiple Sclerosis: A UK Biobank Study. *Neurol. Neuroimmunol. neuroinflammation* **2021**, *8*, doi:10.1212/NXI.0000000000001007.

131. Almasy, L.; Towne, B.; Peterson, C.; Blangero, J. Detecting genotype x age interaction. *Genet. Epidemiol.* **2001**, *21 Suppl 1*, doi:10.1002/GEPI.2001.21.S1.S819.

132. Poveda, A.; Chen, Y.; Brändström, A.; Engberg, E.; Hallmans, G.; Johansson, I.; Renström, F.; Kurbasic, A.; Franks, P.W. The heritable basis of gene-environment interactions in cardiometabolic traits. *Diabetologia* **2017**, *60*, 442–452, doi:10.1007/S00125-016-4184-0.

133. Hujoel, M.L.A.; Gazal, S.; Loh, P.R.; Patterson, N.; Price, A.L. Liability threshold modeling of case–control status and family history of disease increases association power. *Nat. Genet. 2020 525* **2020**, *52*, 541–547, doi:10.1038/s41588-020-0613-6.

134. Elston, R.C.; Olson, J.M.; Palmer, L. Biostatistical genetics and genetic epidemiology. **2002**, 831.

135. Kraft, P.; Thomas, D.C. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am. J. Hum. Genet.* **2000**, *66*, 1119–1131, doi:10.1086/302808.

136. Carpenter, B.; Gelman, A.; Hoffman, M.D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.A.; Guo, J.; Li, P.; Riddell, A. Stan: A Probabilistic Programming Language. *J. Stat. Softw.* **2017**, *76*, 1–32, doi:10.18637/JSS.V076.I01.

137. Jonah Gabry and Rok Cesnovar cmdstanr: R Interface to "CmdStan." **2021**.

138. Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. **2018**.

139. Stan Development Team Stan User's Guide, 2.31. **2022**.

140. Montomoli, C.; Allemani, C.; Solinas, G.; Motta, G.; Bernardinelli, L.; Clemente, S.; Murgia, B.S.; Ticca, A.F.; Musu, L.; Piras, M.L.; et al. An ecologic study of geographical variation in multiple sclerosis risk in central Sardinia, Italy. *Neuroepidemiology* **2002**, *21*, 187–193, doi:10.1159/000059522.

141. Ebers, G.C.; Sadovnick, A.D.; Dyment, D.A.; Yee, I.M.L.; Willer, C.J.; Risch, N. Parent-of-origin effect in multiple sclerosis: Observations in half-siblings. *Lancet* **2004**, *363*, 1773–1774, doi:10.1016/S0140-6736(04)16304-6.

142. Hoppenbrouwers, I.A.; Liu, F.; Aulchenko, Y.S.; Ebers, G.C.; Oostra, B.A.; Van Duijn, C.M.; Hintzen, R.Q. Maternal transmission of multiple sclerosis in a dutch population. *Arch. Neurol.* **2008**, *65*, 345–348, doi:10.1001/ARCHNEUROL.2007.63.

143. Kendler, K.S.; Ohlsson, H.; Lichtenstein, P.; Sundquist, J.; Sundquist, K. The Nature of the Shared Environment. *Behav. Genet.* **2019**, *49*, 1–10, doi:10.1007/S10519-018-9940-0.

144. Nakagawa, S.; Schielzeth, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **2013**, *4*, 133–142, doi:10.1111/J.2041-210X.2012.00261.X/FULL.

145. Steri, M.; Orrù, V.; Idda, M.L.; Pitzalis, M.; Pala, M.; Zara, I.; Sidore, C.; Faà, V.; Floris, M.; Deiana, M.; et al. Overexpression of the Cytokine BAFF and Autoimmunity Risk. *N. Engl. J. Med.* **2017**, *376*, 1615–1626, doi:10.1056/NEJMOA1610528.

146. Waubant, E.; Lucas, R.; Mowry, E.; Graves, J.; Olsson, T.; Alfredsson, L.; Langer-Gould, A. Environmental and genetic risk factors for MS: an integrated review. *Ann. Clin. Transl. Neurol.* **2019**, *6*, 1905–1922, doi:10.1002/ACN3.50862.

147. Mameli, G.; Cossu, D.; Cocco, E.; Masala, S.; Frau, J.; Marrosu, M.G.; Sechi, L.A. EBNA-1 IgG titers in Sardinian multiple sclerosis patients and controls. *J. Neuroimmunol.* **2013**, *264*, 120–122, doi:10.1016/J.JNEUROIM.2013.07.017.

148. Alfredsson, L.; Olsson, T. Lifestyle and Environmental Factors in Multiple Sclerosis. *Cold Spring Harb. Perspect. Med.* **2019**, *9*, doi:10.1101/CSHPERSPECT.A028944.

149. Ascherio, A.; Munger, K.L.; Lünemann, J.D. The initiation and prevention of multiple sclerosis. *Nat. Rev. Neurol.* **2012**, *8*, 602–612, doi:10.1038/NRNEUROL.2012.198.

150. Amato, M.P.; Derfuss, T.; Hemmer, B.; Liblau, R.; Montalban, X.; Soelberg Sørensen, P.; Miller, D.H.; Alfredsson, L.; Aloisi, F.; Ascherio, A.; et al. Environmental modifiable risk factors for multiple sclerosis: Report from the 2016 ECTRIMS focused workshop. *Mult. Scler.* **2018**, *24*, 590–603, doi:10.1177/1352458516686847.

151. Bediako, Y.; Adams, R.; Reid, A.J.; Valletta, J.J.; Ndungu, F.M.; Sodenkamp, J.; Mwacharo, J.; Ngoi, J.M.; Kimani, D.; Kai, O.; et al. Repeated clinical malaria episodes are associated with modification of the immune system in children. *BMC Med.* **2019**, *17*, doi:10.1186/s12916-019-1292-y.

152. Natama, H.M.; Rovira-Vallbona, E.; Krit, M.; Guetens, P.; Sorgho, H.; Somé, M.A.; Traoré-Coulibaly, M.; Valéa, I.; Mens, P.F.; Schallig, H.D.F.H.; et al. Genetic variation in the immune system and malaria susceptibility in infants: a nested case–control study in Nanoro, Burkina Faso. *Malar. J.* **2021**, *20*, doi:10.1186/s12936-021-03628-y.

153. Garn, H.; Potaczek, D.P.; Pfefferle, P.I. The Hygiene Hypothesis and New Perspectives—Current Challenges Meeting an Old Postulate. *Front. Immunol.* 2021, *12*, doi:10.3389/fimmu.2021.637087.

154. Arama, C.; Quin, J.E.; Kouriba, B.; Farrants, A.K.Ö.; Troye-Blomberg, M.; Doumbo, O.K. Epigenetics and Malaria Susceptibility/Protection: A Missing Piece of the Puzzle. *Front. Immunol.* **2018**, *9*, doi:10.3389/FIMMU.2018.01733.

155. Mazzone, R.; Zwergel, C.; Artico, M.; Taurone, S.; Ralli, M.; Greco, A.; Mai, A. The emerging role of epigenetics in human autoimmune disorders. *Clin. Epigenetics 2019 111* **2019**, *11*, 1–15, doi:10.1186/S13148-019-0632-2.

156. Küçükali, C.İ.; Kürtüncü, M.; Çoban, A.; Çebi, M.; Tüzün, E. Epigenetics of multiple sclerosis: an updated review. *Neuromolecular Med.* **2015**, *17*, 83–96, doi:10.1007/S12017-014-8298-6.

157. Puthenparampil, M.; Perini, P.; Bergamaschi, R.; Capobianco, M.; Filippi, M.; Gallo, P. Multiple sclerosis epidemiological trends in Italy highlight the environmental risk factors. *J. Neurol.* **2022**, *269*, 1817–1824, doi:10.1007/S00415-021-10782-5/FIGURES/3.

158. Nova, A.; Fazia, T.; Beecham, A.; Saddi, V.; Piras, M.; McCauley, J.L.; Berzuini, C.; Bernardinelli, L. Plasma Protein Levels Analysis in Multiple Sclerosis Sardinian Families Identified C9 and CYP24A1 as Candidate Biomarkers. *Life (Basel, Switzerland)* **2022**, *12*, doi:10.3390/LIFE12020151.

159. Nova, A.; Baldrighi, G.N.; Fazia, T.; Graziano, F.; Saddi, V.; Piras, M.; Beecham, A.; McCauley, J.L.; Bernardinelli, L. Heritability Estimation of Multiple Sclerosis Related Plasma Protein Levels in Sardinian Families with Immunochip Genotyping Data. *Life (Basel, Switzerland)* **2022**, *12*, doi:10.3390/LIFE12071101.

160. Frau, J.; Coghe, G.; Lorefice, L.; Fenu, G.; Cocco, E. Infections and Multiple Sclerosis: From the World to Sardinia, From Sardinia to the World. *Front. Immunol.* **2021**, *12*, 4143, doi:10.3389/FIMMU.2021.728677/BIBTEX.

161. Kumar, A.; Cocco, E.; Atzori, L.; Marrosu, M.G.; Pieroni, E. Structural and Dynamical Insights on HLA-DR2 Complexes That Confer Susceptibility to Multiple Sclerosis in Sardinia: A Molecular Dynamics Simulation Study. *PLoS One* **2013**, *8*, e59711, doi:10.1371/JOURNAL.PONE.0059711.

Acknowledgements

Thanks to my family for supporting me in my choices and giving me the opportunity to study all these years.

A special thanks to Prof. Luisa Bernardinelli and Dott. Teresa Fazia for giving me the opportunity to work as a researcher, and for their kindness in helping me when needed.

Thanks to my colleagues for their friendships and time spent together.