University of Eastern Piedmont "Amedeo Avogadro"

School of Medicine

Department of Health Sciences

Master's degree course in Medical Biotechnologies

Master Thesis

COMPARATIVE ANALYSIS OF CLINICAL PROGRESSION PHENOTYPES WITH GENETIC VARIANTS AND ENVIRONMENTAL FACTORS IN MULTIPLE SCLEROSIS

Supervisor:

Prof.ssa Nadia Barizzone

Candidate: Sofia Terreno

Matricula Number: 20037603

Academic year: 2025/2026

Table of Contents

Summary

Multiple sclerosis (MS) is a chronic autoimmune disease of the central nervous system characterized by substantial clinical heterogeneity, with disability accumulating through relapse-associated worsening (RAW) and progression independent of relapse activity (PIRA). Although genome-wide association studies (GWAS) have identified more than 230 genetic variants associated with MS susceptibility, the genetic determinants of disease progression remain poorly understood. In addition, environmental factors such as cigarette smoking may contribute to long-term disability accumulation.

This study investigated the relationship between selected genetic variants, polygenic risk scores (PRSs), and cigarette smoking with disability progression in approximately 1,500 patients with MS enrolled in a pilot linkage project between the PROGEMUS Consortium and the Italian Multiple Sclerosis Registry (RISM). Environmental data were obtained through the EnviMS questionnaire, while genetic data have been obtained through GWAS analysis, implemented by SNP and HLA allele imputation, single marker genotyping and calculation of 7 PRSs. Associations with clinical outcomes were evaluated using statistical and survival analyses.

HLA-A*02:01, traditionally considered a protective allele for MS susceptibility, showed an unexpected association with PIRA, suggesting that genetic determinants of susceptibility and progression may differ. Instead, HLA-DRB1*15:01, a major susceptibility allele, showed limited prognostic value for disability progression. APOE ε4 was not associated with progression outcomes, whereas the disease severity-associated variant rs10191329 was associated with RAW. Susceptibility-based PRSs showed limited predictive ability, while the severity-specific PRS was more strongly associated with disability accumulation. Cigarette smoking was also associated with worse clinical outcomes, highlighting its role as an important environmental factor predictor of disability progression.

Overall, these findings suggest that genetic architecture underlying MS susceptibility only partially overlaps with that governing disease progression. Progression-specific genetic markers and severity-based polygenic scores may provide greater prognostic value than traditional susceptibility markers, while smoking further emphasizes the importance of modifiable environmental factors. Although these findings require further validation, they contribute to a better understanding of the complex interplay between genetics, environmental exposures, and clinical progression in MS, supporting future personalized prognostic strategies.

Introduction

1. MULTIPLE SCLEROSIS

1.1. Definition and clinical relevance

Multiple sclerosis (MS) is a complex autoimmune disorder of the central nervous system (CNS), characterized by demyelination and progressive neurodegeneration[1]. MS represents one of the leading causes of non-traumatic neurological disability in young adults, with substantial consequences for physical function, cognitive performance, quality of life and socioeconomic burden[2]; the number of confirmed MS cases worldwide is estimated to be approximately 2.9 million, of which around 126,000 are reported in Italy[3]. First described in 1868 by the French neurologist Jean Martin Charcot, MS has been intensively studied over the last decades, leading to a better understanding of its pathogenetic mechanisms and, consequently, to the development of new therapies and diagnostic methods[4]. MS is classified as a multifactorial disease, meaning that its development is influenced by the interaction between lifestyle or environmental risk factors and genetic predisposition. As a complex disease, MS is affected by thousands of common genetic variants, each exerting an individual effect on disease susceptibility and progression. This characteristic makes the study of such conditions particularly challenging due to their complex genetic architecture[5].

1.2. Epidemiology and geographical distribution

Like most autoimmune disorders, MS affects women more frequently, with a female-to-male ratio of 3:1 and typically has an onset between 20 years and 40 years of age. However, MS can develop at any age: approximately 10% of cases are diagnosed before the age of 18, whereas cases diagnosed after the age of 50 are referred as late-onset multiple sclerosis (LOMS), which represents a relatively rare form of the disease[6]. From a geographical distribution perspective, the prevalence of multiple sclerosis varies across countries and continents, exhibiting a latitude-dependent gradient. In Europe, for instance, individuals living in regions farther form the Equator show a higher risk of developing the disease compared with those living in the southern parts of the continent. A similar trend can be observed also when analysing the incidence of MS[7].

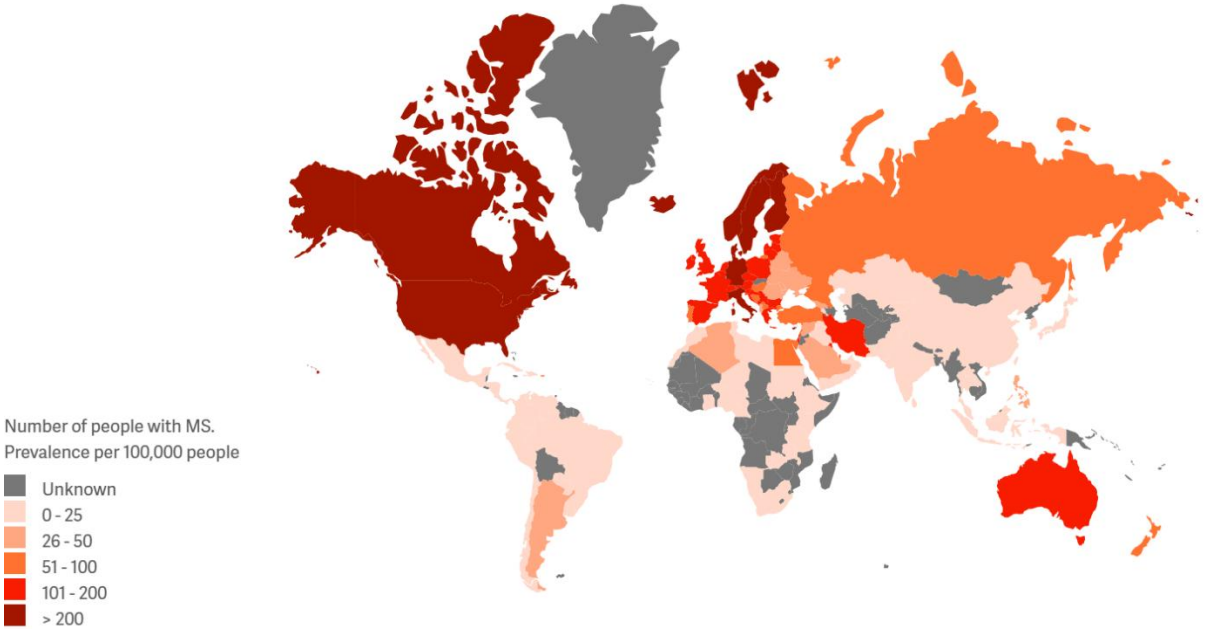

Figure 1. Worldwide geographical distribution of MS in 2026. This choropleth map illustrates the worldwide distribution and prevalence rates of individuals living with MS, highlighting a higher concentration of cases in northern latitudes (e.g., North America and Northern/Central Europe) and Australia. Globally, an estimated 2.9 million people are living with the condition. Data categories range from unknown (grey) to a prevalence exceeding 200 cases per 100,000 individuals (dark red)[8].

2. CLINICAL FEATURES OF MS

2.1. Main clinical phenotypes

The first official classification of MS phenotypes was carried out by the U.S. National Multiple Sclerosis Society (NMSS) in 1996 with the purpose of creating a standardized terminology in the clinical field. The phenotypes identified were: Relapsing Remitting MS (RRMS), Primary Progressive MS (PPMS), Secondary Progressive MS (SPMS) and Progressive Relapsing MS (PRMS) – now eliminated[8]. With the 2012 revision of the NMSS classification, the description of the different subtypes was refined including a new term: Clinacally Isolated Syndrome (CIS)[9].

Found in about 85% of patients, RRMS is considered the most prevalent phenotype; it is characterized by periods of clinical stability without neurological symptoms – remissions and periods of neurological disfunction – relapses, with a frequency of relapses that differs from patient to patient. SPMS is the evolution of RRMS, and it consists of a gradual progression of

the disease with occasional events of relapse. PPMS affects about 10-20% of patients and differs from the SPMS because of the absence of an initial phase of RRMS. Lastly, CIS is defined as the first neurological episode suggestive of MS, but without the dissemination in time[10].

Disability accumulation in multiple sclerosis may occur through different mechanisms, not necessarily related to clinical relapses. In particular, Progression Independent of Relapse Activity (PIRA) refers to the gradual worsening of disability in the absence of relapse activity, whereas Relapse-Associated Worsening (RAW) results from incomplete recovery following a relapse event. These two events are considered contributors to Confirmed Disability Accumulation (CDA), a measure of sustained disability worsening over time that can be estimated using the Expanded Disability Status Scale (EDSS) [11]. Importantly, PIRA has been observed not only in progressive forms of MS but also in patients with CSI and RRMS, suggesting that progression independent of relapses may be present from the earliest stage of the disease[12].

2.2. Diagnosis and McDonald criteria

Multiple sclerosis is a multifactorial disease that does not possess a specific biological biomarker; for this reason, the McDonald diagnostic criteria for multiple sclerosis were established in 2001 to help healthcare providers accurately diagnose the pathology. According to these criteria, the diagnosis of MS is based on a combination of clinical examination, laboratory analyses and magnetic resonance imaging (MRI); since 2001, McDonald criteria have been repeatedly updated, most recently in 2024[13]. Following the McDonald criteria guidelines, after excluding the presence of any other condition presenting multiple sclerosis-like symptoms, dissemination in space (DIS) must be identified to establish the diagnosis of MS. DIS is defined when 2 or more of the 5 CNS anatomical sites show typical lesions. In addition to DIS, another parameter that can reconfirm the diagnosis of MS is dissemination in time (DIT); DIT is useful to establish the continuousness of the disease, with the presence of episodes distributed over time, rather than a single isolated episode[14].

3. PATHOPHYSIOLOGY OF MS

The neuropathogenesis of multiple sclerosis is a complex process involving several mechanisms including demyelination, immune system (IS) activation and neurodegeneration. Demyelination is a defining feature of multiple sclerosis and occurs as a result of an immune system attack against oligodendrocytes, leading to the disruption of myelin sheaths on axons.

This process ultimately contributes to plaque formation and impaired nerve conduction, particularly in the white matter regions of the brain, optic nerve, and spinal cord. In the acute phase of MS, this process is associated with infiltration of the tissue by macrophages and lymphocytes; when the inflammation persists (in approximately 20% of lesions), the inflammatory process becomes more organized, with accumulation of memory T cells and monocytes. A regenerative process is initiated by oligodendrocyte progenitor cells (OPCs), which attempt to repair myelin sheaths; however, under chronic condition, OPCs fail to fully differentiate and are unable to effectively complete the regeneration process[15]. The absence of myelin sheaths leads to an abnormal distribution of ion channels along axons; this induces an ionic imbalance within the axon, resulting in calcium accumulation that stimulates axonal proteolysis[16].

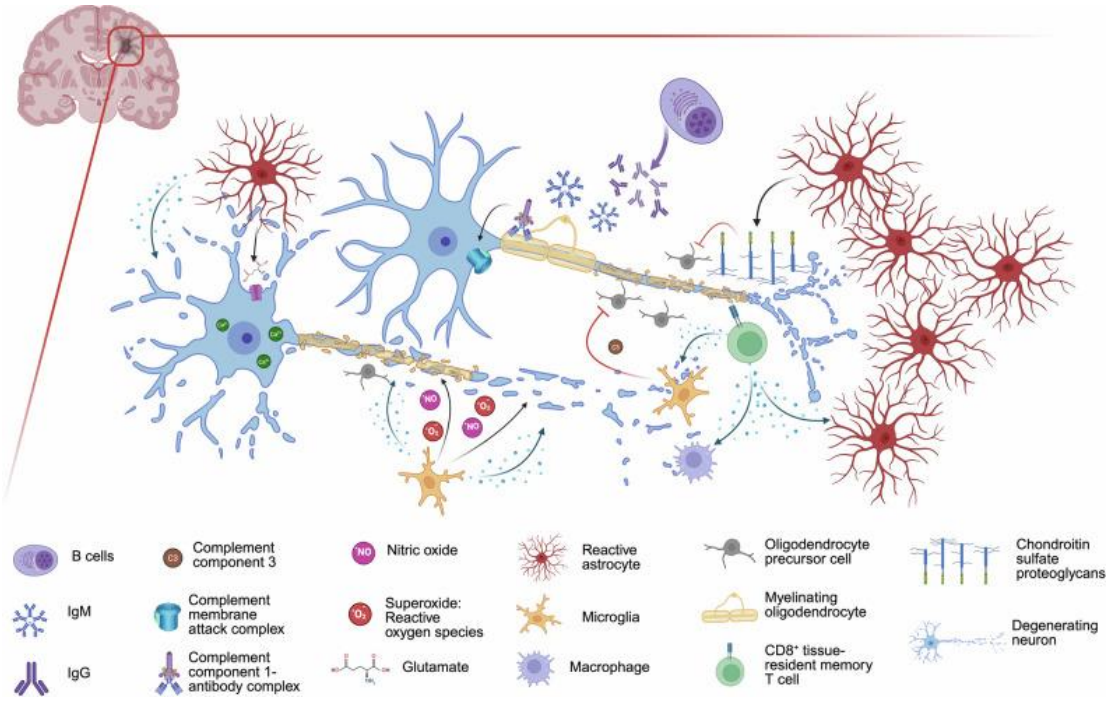

Figure 2. Pathological mechanisms of neurodegeneration and demyelination in MS. Schematic overview of a localized CNS lesion depicting the multicellular interactions driving demyelination and neuronal injury. The illustration details antibody-mediated damage via B cells secreting IgM and IgG, complement system activation, and T-cell-mediated responses. The right portion highlights the cellular response to injury[15].

4. IMMUNOPATHOLOGY OF MS

Chronic inflammation is recognized as a hallmark of neurodegenerative disease, including MS; autoreactive lymphocytes – primarily T and B cells – directed against myelin are able to cross the blood-brain barrier (BBB) and migrate into the CNS, triggering an inflammatory cascade that leads to demyelination. Following myelin breakdown, immune cells can further infiltrate nervous tissue, promoting a state of chronic inflammation[17].

4.1. Autoreactive T cells

MS is generally considered a T cell-mediated disease due to the association between MS susceptibility and genetic variants related to pathogenic T cell differentiation, including interleukin 2 (IL-2) and the IL-7 receptor (IL-7R). Moreover, mutations in the major histocompatibility complex class II (MHCII) genes – whose protein present antigen to the T cell receptor (TCR) – can increase an individual's susceptibility to MS[18]. Several proteins located within the myelin sheath – including myelin basic protein (MBP), myelin-associated glycoprotein (MAG) and myelin oligodendrocyte glycoprotein (MOG) – have been identified as target of autoreactive lymphocytes, with MBP appearing to be the most immunogenic[19]. CD4+ T cells – especially Th1 and Th17 – are essential for initiating and maintaining the autoimmune response; however multiple studies have also demonstrated the crucial role of CD8+ T cells and macrophages. Overactivation of Th1 and Th17 cells leads to an accumulation of interleukins and cytokines, that create a state of inflammation in the surrounding tissue. In addition to the autoreactive activity exerted by lymphocytes, the development of MS is also characterized by reduced activity of regulatory T cells. Together, these cells are able to cause the death of myelin-producing oligodendrocytes, damage the myelin sheath surrounding nerve fibres, and compromise brain tissue integrity, ultimately leading to lesion formation within the CNS[20].

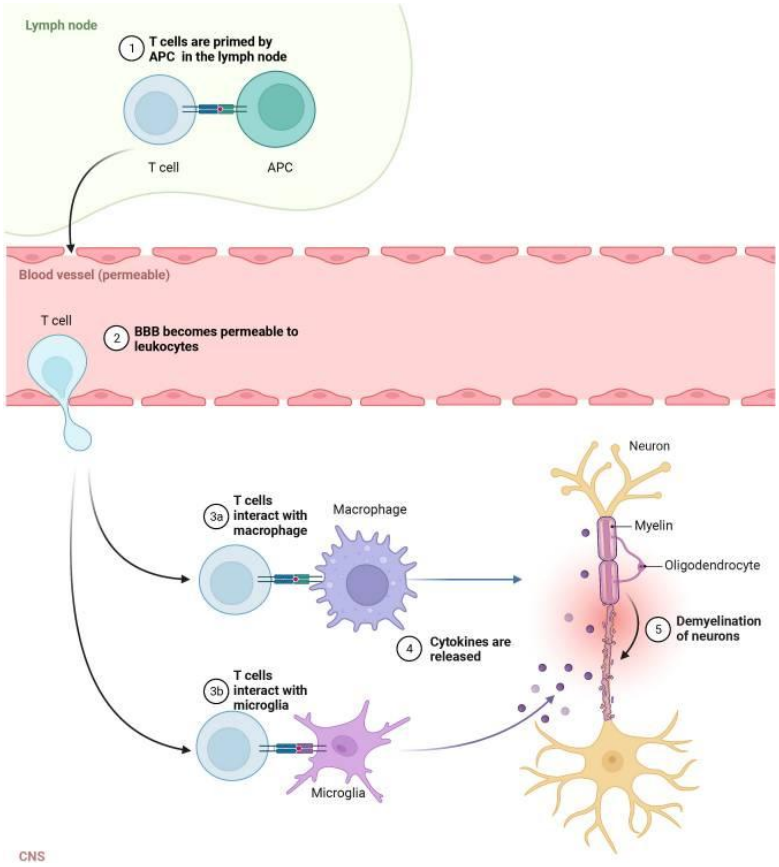

Figure 3. Mechanism of T-cell-mediated extravasation and subsequent demyelination within the CNS. This schematic illustration outlines the chronological sequence of immune-mediated nerve damage. (1) Initial priming of T cells by APCs within the peripheral lymph node. (2) Migration of activated T cells to a permeable blood vessel, where they traverse the blood-brain barrier (BBB). (3a, 3b) Upon entering the central nervous system (CNS), T cells interact with localized macrophages and microglia. (4) This interaction triggers the release of pro-inflammatory cytokines, culminating in (5) the destruction of the myelin sheath surrounding the axon (demyelination), disrupting the functional connection between the oligodendrocyte and the neuron[19].

4.2. Autoreactive B cells

Multiple sclerosis was originally considered a predominantly T cell-mediated disease, mainly characterized by an imbalance between pro-inflammatory and anti-inflammatory immune responses within the CNS. However, increasing evidence has demonstrated that B cells also play a crucial role in the pathogenesis of the disease. B cells are now recognized as important contributors to MS pathology across multiple phases and aspects of disease progression, through both antibody-dependent and antibody-independent mechanisms[21]. Their functions include antigen presentation, cytokine secretion and the production of autoantibodies that may contribute to CNS damage. In MS lesions, the presence of co-localized immunoglobulin (Ig) and complement deposition in areas of active demyelination strongly supports the involvement

of autoantibodies in the destruction of CNS tissue. These findings suggest that humoral immunity plays a significant role alongside T cell-mediated pathogenesis. The wide variety of autoantibodies detect in patients with MS likely reflects the multifactorial and heterogeneous nature of the disease, as well as the presence of different underlying pathogenic mechanisms among individuals[22].

5. ETIOLOGY OF MS

5.1. Environmental risk factors

Multiple sclerosis is a multifactorial disease in which genetic susceptibility alone is not sufficient to explain disease onset. Environmental and lifestyle-related factors play a major role in modulating both the risk of developing MS and the clinical course of the disease. Several of these factors act through immune system dysregulation and may interact with genetic predisposition. Among the most consistently recognized environmental contributors are vitamin D deficiency, smoking, Epstein–Barr virus (EBV) infection, obesity, and alterations of the gut microbiome[23].

5.1.1. Vitamin D and sun exposure

Low vitamin D status is considered one of the best-established environmental risk factors for MS. Since most vitamin D is produced through skin exposure to ultraviolet B (UVB) radiation rather than dietary intake, reduced sun exposure has been strongly associated with increased MS risk[24]. This could partially explain the geographical distribution of MS as countries located farther from the Equator – where sun exposure is lower – generally show a higher incidence of the disease. Vitamin D is thought to influence both innate and adaptive immune responses by reducing pro-inflammatory activity and promoting immune tolerance. It can modulate dendritic cell maturation, decrease pro-inflammatory cytokine production, and affect antigen presentation, thereby contributing to immune homeostasis[25]. Lower serum vitamin D levels have also been associated with increased disease activity in MS patients, suggesting a role not only in disease susceptibility but also in relapse frequency and progression. Although vitamin D supplementation studies have not always shown statistically significant primary outcomes, several findings support its potential beneficial role in disease management[26]. The action of vitamin D in the human body is mediated through the VD receptor (VDR), a nuclear transcription factor that regulates the expression of several genes in the human genome[27]. Different polymorphisms of VDR gene have been identified, each of which can influence the

binding efficiency of vitamin D to its receptor. Therefore, several studies have shown that certain VDR polymorphisms may be associated with an increased risk of MS[28].

5.1.2. Smoking

Cigarette smoking is a well-established environmental risk factor for MS and has been associated with both increased disease susceptibility and worse disease progression. Several studies investigating the correlation between MS and smoking have shown that individuals who actively smoke are approximately 50% more likely to be diagnosed with MS than non-smokers. A similar trend has also been observed in disease progression, as smoking may accelerate the progression of MS. Negative effects on MS can also be induced by passive smoking, as demonstrated by several studies investigating the relationship between parental smoking and MS in children[29]. In general, smoking is known to influence MS development mainly by enhancing the inflammatory response, increasing the levels of pro-inflammatory cytokines, and promoting oxidative stress. Smoking affects T cells, B cells, and other immune cells, while nicotine suppresses T cell responses and leads to dysfunction of antigen-presenting cells (APC)[30]. Moreover, smoking influences the presence of antigen in the blood, as it can regulate antigen release through tissue hypoxia or cellular necrosis. Overall, smoking promotes systemic inflammation, oxidative stress, and immune dysregulation, all of which may contribute to CNS autoimmunity[31].

5.1.3. Epstein-Barr virus (EBV) infection

Among infectious factors, Epstein–Barr virus (EBV) is currently considered one of the strongest candidates involved in MS pathogenesis and may represent a leading causal factor[24].

EBV is a highly prevalent herpesvirus that establishes lifelong latent infection, primarily in B cells, in more than 90% of adults worldwide. It was the first human tumour virus ever identified and, since its discovery more than 50 years ago, has been detected in numerous cancers; moreover, it is the most common cause of infectious mononucleosis. EBV is a herpesvirus with a double-stranded DNA genome, that is able to persists in the nucleus of infected B cells in a circular chromatin structure known as “episome”. After an initial phase in which EBV infects squamous epithelial cells, it can cross the mucosa and reach B lymphocytes in the tonsils; infection of B cells initiate a process of cellular reprogramming that leads to the generation of memory B cells harbouring EBV genome[32]. EBV life cycle alternates between latent and lytic phase. During the latency phase, the virus expresses only a specific set of genes and relies on

the host DNA replication machinery, enabling long-term survival of infected cells without triggering a strong immune response. Under certain conditions, EBV can enter the lytic cycle, during which all viral genes are expressed and new virions are produced. Several latency-associated genes, including EBNA1, EBNA2, EBNA3A, EBNA3C, and LMP1, are essential for B cell immortalization[33].

Increasing epidemiological and mechanistic evidence supports EBV role in triggering MS in genetically susceptible individuals. Infection with EBV is nearly universal in adults, and individuals who developed symptomatic infectious mononucleosis show a significantly increased risk of developing MS in later life. A meta-analysis has demonstrated that individuals with a history of clinically overt infectious mononucleosis have more than a twofold increased risk of developing MS compared to those without such history[34]. The mechanisms underlying this association involve complex interactions between the virus and the host immune system, especially within the B-cell compartment.

Epstein-Barr nuclear antigen 1 (EBNA1) plays a key role in viral latency, B-cell immortalization, and neoplastic growth. This protein is required for the maintenance of the episome within the host cell nucleus and for its replication through the host replication machinery. It ensures that the viral genome is properly replicated and segregated during host cell division by interacting with host DNA replication machinery. In addition to its role in viral persistence, EBNA1 contributes to immune evasion. It has been shown to interfere with host immune pathways, thereby reducing the ability of the IS to detect and eliminate infected cells[35,36]. This property supports long-term viral latency and may contribute to sustained immune activation observed in MS.

Epstein-Barr nuclear antigen 2 (EBNA2) is another critical latency-associated protein that plays a central role in regulating viral and host gene expression. EBNA2 acts as a transcriptional activator, influencing the expression of both viral genes and host cellular genes involved in B-cell activation and proliferation[36]. Through its interaction with host transcriptional machinery, EBNA2 drives the reprogramming of infected B cells, promoting their proliferation and survival. This activity is essential for the establishment of latent infection and the generation of long-lived infected memory B cell[35]. Importantly, EBNA2 has been linked to genetic susceptibility loci associated with MS. It can interact with regions of the human genome involved in immune regulation, suggesting a potential mechanism by which EBV infection may influence disease risk through modulation of host gene expression[32,37].

The contribution of EBV to MS pathogenesis is thought to involve multiple, non-mutually exclusive mechanisms, including molecular mimicry, persistence of autoreactive B cells, viral infiltration of CNS, impaired cytotoxic T cell control of EBV infection, inflammation, dysregulation of B-cell gene expression, and interaction with HLA[32]. Several EBV antigens have been identified as target of autoantibodies detected in MS through peptide library analyses; for example, a domain of EBNA1 mimics glial cell adhesion molecule[32,38]. Moreover, antibody reactivity to EBNA1 appears to be higher in individuals carrying HLA DRB1*1501 allele, suggesting that the HLA DRB1*1501 allele and EBNA1-specific antibodies may act synergically as risk factors for MS[39]. The interaction between EBV infection and HLA-associated genetic risk further supports the hypothesis that adaptive immunity plays a central role in MS development. Although EBV infection alone is not sufficient to cause MS, it is considered a major contributor to disease initiation.

5.1.4. Body Mass Index (BMI) and obesity

Adipose tissue is hormonally active and secretes several molecules involved in the inflammatory processes, such as cytokines and adipokines. Condition such as obesity can alter the adipokine profile, shifting it towards a more pro-inflammatory state, which may favour the development of autoimmune responses. Elevated body mass index (BMI) and obesity play a key role in MS development; moreover, it has been observed that obesity can also influence the level of clinical disability in MS patients at the time of diagnosis and during disease progression[40]. Similar to smoking and EBV infection, obesity appears to interact with HLA susceptibility genes, amplifying disease risk in predisposed individuals[41].

5.1.5. Gut microbiota alterations

The intestinal microbiota plays an important role in immune regulation, and alterations in microbial composition have been associated with several inflammatory conditions. Comparative studies between MS patients and healthy controls have revealed significant differences in gut microbial profiles, suggesting an association between microbiome composition and both disease risk and clinical course[42]. A study comparing the gut microbiome of MS patients and healthy controls reported alterations that may be linked to changes in the immune transcriptome and to treatment effects[43]. Overall, these findings support the hypothesis that gut microbiome dysbiosis may contribute to MS pathogenesis through modulation of immune responses.

5.2. Genetic risk factors

Multiple sclerosis is widely recognized as a multifactorial disease in which genetic susceptibility plays a crucial role alongside environmental and lifestyle factors. Family and population studies have consistently demonstrated that individuals with affected relatives have a higher risk of developing MS, supporting a significant heritable component[44]. However, the genetic architecture of MS is highly polygenic, meaning that disease susceptibility is influenced by the combined effect of numerous genetic variants, each contributing modestly to overall risk[45].

The first genetic locus associated with MS susceptibility was identified in the 1970s; this finding was later redefined to the human leukocyte antigen (HLA) gene cluster within the highly polymorphic major histocompatibility complex (MHC) region on chromosome 6, and more specifically to the HLA-DRB1*15:01 allele, which represent the strongest genetic risk factor for MS[46]. Large-scale genomic studies, particularly genome-wide association studies (GWAS), have substantially expanded the understanding of MS genetics by identifying a wide range of susceptibility loci, many of which are involved in immune system regulation[47]. To date, more than 230 genetic variants regarding disease susceptibility have been identified through this type of analysis.

5.2.1. Role of the Major Histocompatibility Complex (MHC)

The strongest genetic association with MS susceptibility is located within the major histocompatibility complex (MHC) on chromosome 6p21. This region contains genes – HLA class I, II and III – encoding polymorphic cell surface glycoproteins, which are essential for adaptive immune responses. Among these, the HLA class II region (HLA-DR2) has been consistently identified as the most significant genetic determinant of MS risk. MHC proteins play a crucial role in the immune response, as they present antigens to immune cell, enabling the initiation of immune signalling. They are essential for distinguishing between self and non-self-antigens, and defects in their function can lead to immune system-related diseases. Variants within this region influence how antigens are presented to T cells, thereby shaping immune responses and potentially promoting autoimmunity[48]. In general, ligands presented by MHC molecules are protein-derived peptides that undergo proteolytic processing to become suitable for antigen presentation. Typically, MHC class I and class II molecules present antigenic peptides to $CD8^+$ T cells and $CD4^+$ T cells, respectively[49].

5.2.2. HLA-DRB1*15:01 as the Major Risk Allele

The HLA-DRB1*15:01 allele represents the strongest and most reproducible genetic risk factor for MS identified to date. Individuals carrying HLA-DRB1*15:01 have a significantly increased risk of developing MS compared to non-carriers[45]. The association between HLA-DRB1*15:01 and MS follows a dose-response model; the effect of this allele does not follow an on-off mechanism but rather depends on the number of copies of the allele present in an individual, a pattern that has also been observed in other autoimmune diseases. Consequently, homozygous individuals exhibit a higher risk that heterozygous carriers. Interestingly, DRB1*15:01 is found at relatively high frequencies in European and Asian populations, whereas it is relatively rare in other populations; in African Americans, the major risk variant has been identified HLA-DRB1*15:03[48].

In addition to the effect of the HLA-DRB1*15:01 risk allele alone, its impact has been shown to increase further when it is present in a heterozygous state together with the HLA-DRB1*08:01 allele, whereas DRB1*08:01 alone does not appear to significantly influence MS risk[50]. Besides *15:01, several other HLA-DBR1 variants, including *03:01, *13:03, *04:04, *04:01, and *14:01, have been independently associated with MS susceptibility. Most of these association are linked to four amino acid changes in different position of the HLA-DR molecules. Since these amino acids are located within the peptide-binding groove, they can influence antigen binding and recognition by immune cells[46].

5.2.3. HLA-A*02:01 as protective genetic factor

In contrast with HLA-DRB1*15:01, which confers the strongest genetic susceptibility to MS, the HLA-A*02:01 is a class I HLA allele has consistently been associated with a reduced risk of disease development. HLA-A*02:01 is a class I HLA allele whose protective effect appears to be independent of the influence exerted by HLA-DRB1*15:01. Large genetic association studies have demonstrated that individuals carrying HLA-A*02:01 exhibit a significantly lower likelihood of developing MS compared with non-carriers. Furthermore, the protective effect of HLA-A*02:01 may partially counterbalance the increased susceptibility conferred by HLA-DRB1*15:01, highlighting the complex interplay between class I and II HLA molecules in shaping MS risk[51].

5.2.4. Non-HLA Genetic Variants associated with MS susceptibility

Although the MHC region represent a major component of genetic susceptibility to MS, several non-HLA loci have been associated with disease risk. Most of these variants involve genes related to immune system regulation, particularly pathways controlling cytokine signalling, lymphocyte activation, and immune cell differentiation. Examples include genes encoding interleukins and their receptors, as well as molecules involved in T-cell receptor signalling pathways. Together, these genetic variants can alter the balance between pro-inflammatory and anti-inflammatory immune responses, thereby contributing to MS pathogenesis[48]. In the largest genome-wide association study conducted to date, the International Multiple Sclerosis Genetics Consortium (IMSGC) identified 201 non-HLA susceptibility loci (200 autosomal loci and one locus on chromosome X), together with 32 independent variants within the extended HLA region. In addition, the study reported more than 500 suggestive susceptibility loci, providing a comprehensive genetic map of MS and highlighting the predominant role of immune-related pathways in disease susceptibility[52].

The first MS-associated genetic variant located outside the MHC region was identified in 2007. This variant, rs12722489, is positioned within the first intron of the IL2RA gene, which encodes the alpha chain of the interleukin-2 receptor, a molecule involved in multiple immune regulatory pathways. To date, several loci outside the MHC gene have been identified, including a chromosome X variant associated with MS. The associated variant, rs2807267, is located within T cell-specific enhancer region[46].

5.2.5. Genetic Variants associated with disease severity

While most genetic studies have focused on susceptibility to MS, the genetic determinants of disease progression and disability accumulation remain far less understood. To date, only a single genetic variant has reached genome-wide significance for MS severity, highlighting the complexity of identifying genetic factors underlying disease progression. Nevertheless, genome-wide association studies have identified approximately ten additional suggestive severity-associated loci, indicating that further susceptibility variants may emerge as larger and more deeply phenotyped cohorts become available. Overall, the genetic architecture of MS severity is still poorly characterized, and considerably less is known about progression-associated variants than about those influencing disease susceptibility.

A recent genome-wide association study investigating disease progression identified rs10191329 in the DYSF-ZNF638 locus as the first genetic variant significantly associated with MS severity. This risk allele is associated with accelerated disability progression, a shorter time

to requiring a walking aid, and increased cortical and neurodegeneration contribute to long-term disease outcome independently of the immune pathways that drive the disease susceptibility[53].

The apolipoprotein E (APOE) gene, located on chromosome 19, encodes a protein involved in lipid transport, neuronal maintenance, synaptic plasticity, and neuroinflammatory regulation. Three major allelic variants have been identified (ε2, ε3, and ε4), with the ε4 allele representing the strongest common genetic risk factor for late-onset Alzheimer's disease and well-established contributor to neurodegenerative processes. Beyond its established role in neurodegeneration, increasing evidence suggests that APOE may also influence disease progression in MS. In a longitudinal study of 160 patients with relapsing MS, APOE ε4 carriers showed a significantly increased risk of relapse-associated worsening (RAW), with a more than a threefold higher hazard of disability accrual compared with non-carriers. Notably, APOE was associated specifically with disability accumulation linked to inflammatory relapse activity rather than progression independent of relapse activity (PIRA)[54].

Together, these findings suggest that genetic factors may contribute not only to MS susceptibility but also to disease severity and long-term disability accumulation. However, compared with the extensive knowledge of susceptibility loci, the genetic basis of MS progression remains largely unresolved and requires further investigation.

6. GENOME-WIDE ASSOCIATION STUDY (GWAS)

The highly polygenic nature of complex diseases such as multiple sclerosis makes the identification of susceptibility variants particularly challenging. To address this complexity, genome-wide association studies (GWAS) were developed as high-throughput, hypothesis-free approaches capable of analysing genetic variation across the entire. These studies have enabled the identification of numerous genomic loci associated with disease susceptibility, substantially improving the understanding of the genetic architecture underlying complex traits[55].

6.1. Principles of GWAS

A genome-wide association study (GWAS) is a research strategy aimed at identifying genetic variants across the genome that are significantly associated with the risk of developing a disease or exhibiting a specific trait. This approach consists of analysing the genomes of large populations to detect variants that occur more frequently in affected individuals than in

unaffected controls[56]. The most commonly investigated variants in GWAS are single nucleotide polymorphisms (SNPs). A SNP is defined as a variation in a single nucleotide at a specific position in the genome, usually bi-allelic. The frequency of these alleles varies systematically according to the presence, absence or severity of a phenotypic trait. In other words, certain alleles are found more frequently in individuals with a specific disease or characteristic, suggesting a statistical association between that genetic variant and the trait. Is therefore evident that GWAS studies rely on a thorough understanding of the human genome and its architecture, knowledge that has been achieved through major genomic research initiatives aimed at cataloguing human genetic variation. Among these, the 1000 Genome Project, which provided a comprehensive catalogue of both common and rare SNPs across human population[57].

Linkage disequilibrium (LD) refers to the non-random association of alleles at different loci, meaning that different SNPs, usually located close to each other in the genome, tend to be inherited together because of limited recombination during meiosis. LD results in the co-inheritance of multiple variants within a specific locus, making it difficult to identify the true causal variant among them. This is an important concept in human genetics because many genetic associations are influenced by LD. In genetic association studies, the analysis of LD within a candidate region allows the allelic status of neighbouring SNPs to be predicted without directly genotyping all variants[55,58]. This property allows the use of Tag SNPs, which are representative genetic variants that captures the information of surrounding correlated SNPs within a locus, thereby reducing the number of markers that need to be directly genotyped. The degree of LD between two loci can be quantified using the correlation coefficient (r^2), whose values range from 0 to 1. An r^2 value of 0 indicates no correlation between the loci, whereas r^2=1 indicates complete correlation, meaning that the two variants are always inherited together in the studied population.

6.2. Methodology and statistical analysis

Genetic association studies can be broadly classified into two main categories: population-based case-control studies and family based-studies. Family-based approaches are particularly effective for identifying rare variants associated with rare diseases or uncommon sub phenotypes of complex disorders. In contrast, population-based case-control studies are the most widely adopted design for detecting common genetic polymorphisms that contribute to the susceptibility of complex traits and multifactorial diseases, such as MS[59].

6.2.1. Study design and genotyping

Accurate and specific phenotype definition is the first step in the design of case-control studies, as it helps avoid genetic and environmental heterogeneity in underlying causal factors and ensures the replicability of the study, which is a crucial aspect of GWAS[59]. Since the effect size of common variants is generally modest, diseases with lower heritability often require larger sample sizes to achieve adequate statistical power and reliably detect disease-associated loci. Statistical power refers to the probability of detecting a true association between a genetic variant and a phenotype when such an association exists. Statistical power is primarily influenced by sample size, effect size of the association, allele frequency and the significance threshold applied during association testing; it can be estimated using different software tools, such as the Genetic Power Calculator (GPC)[60,61].

Controls should be recruited from the same source population as the cases and should accurately represent individuals who could have developed the disease according to the established case definition; this minimizes the risk of false-positive associations. In genetic association studies, the main source of confounding arises from differences in the ethnic background of cases and controls, a phenomenon commonly referred to as population stratification. Population stratification can influence genetic association studies when two conditions are present: first, disease prevalence differs between the population from which cases and controls are drawn; second, allele frequencies vary across these populations. This may introduce bias because differences in allele frequencies may appear statistically significant due to the sampling scheme rather than to a true effect of the variant on disease risk. Population stratification can be mitigated by matching controls to cases according to ethnicity, and also by matching sex in conditions where disease prevalence differs between males and females[59,62].

Following cohort selection and phenotype definition, genomic DNA is genotyped using high-throughput SNP arrays, enabling the simultaneous analysis of hundreds of thousands of genetic variants across the genome. SNP arrays are a type of DNA microarray technology containing probes specifically designed to target known SNP positions across the genome. Fragmented DNA samples are hybridized to these probes, allowing the identification of the specific alleles present at each SNP locus in the analysed individuals. Owing to their ability to genotype hundreds of thousands of variants simultaneously, SNP arrays have provided a powerful and cost-effective platform for genome-wide studies of complex diseases characterized by intricate patterns of genomic variation[63,64]. Because SNP arrays directly genotype only a subset of

genomic variants, genotype imputation is commonly applied to infer untyped variants using reference panels, such as 1000 Genome Project. Genotype imputation refers to the process of predicting genotypes not directly determined in a sample. To achieve this, imputation methods apply known LD patterns between neighbouring variants, thereby improving the power of association analyses and enabling an indirect broader genome-wide coverage[65].

6.2.2. Quality control procedures

Even the best designed genome-wide association study can be compromised by poor-quality genetic data. For this reason, quality control (QC) procedures represent an essential step in GWAS workflows, ensuring the reliability and reproducibility of downstream association analyses. QC is typically performed at both the sample and the variant level in order to identify technical artefacts, low-quality genotypes, population biases, and potential confounding factors. Individuals or SNPs failing to meet predefined quality thresholds are excluded from subsequent analyses. Common QC criteria include genotype missingness rates, sex discrepancies, heterozygosity rate, relatedness among individuals, deviation from Hardy-Weinberg equilibrium (HWE), minor allele frequency (MAF), and the presence of population outliers[57,66].

Sample-level QC includes:

- Sex discrepancy

Biological sex consistency is assessed by comparing the sex reported in the clinical dataset with the genetically inferred sex based on X chromosome heterozygosity or homozygosity. Male samples are expected to show high X chromosome homozygosity, whereas female typically display lower homozygosity value. Significant discrepancies may indicate sample contamination, mislabelling, or sample swaps. Unless the correct identity of the sample can be verified, individuals showing sex discordance are generally removed from the analysis. In some cases, additional laboratory validation can be performed to clarify the discrepancy, for example through Polymerase Chain Reaction-base assay targeting the Amelogenin gene and the SRY (Sex determining Region Y) gene, which are commonly used for biological sex determination.

- Missingness of genotype

The missingness rate represent the proportion of genotypes that cannot be reliably determined for a given individual. Sample with high missing genotype rates, usually above 3-7%, are commonly excluded because they may reflect poor DNA quality, insufficient DNA quantity, or

technical genotyping failures. A high level of missing data can reduce the reliability of downstream association analyses.

- Relatedness or duplicated

GWAS case-control studies are generally designed under the assumption that analysed individuals are unrelated. Therefore, cryptic relatedness and duplicates samples must be identified and removed to avoid inflation of association signals. Relatedness is commonly evaluated by calculating the identity by descendent (IBD), which estimates the proportion of alleles shared between pairs of individuals. Duplicates usually typically show an IBD value close to 1, whereas first-degree relatives generally present IBD values around 0.5. This analysis is usually performed using only independent SNPs outside the sex chromosomes.

- Ethnic outliers

Population stratification represents one of the main sources of confounding in GWAS, as systematic differences in ancestry between cases and controls my generate false-positive association unrelated to disease biology. For this reason, individuals showing divergent genetic ancestry compared with the main study population are identified and excluded. Population structure is commonly evaluated using principal component analysis (PCA), which allows the detection of ethnic outliers based on genetic similarity patterns[59].

Variant-level QC includes:

- SNP missingness

Variants with hight missing genotype rates across the dataset are excluded, because they may reflect unreliable genotyping performance. Commonly, SNPs with missingness rates exceeding 1-5% are removed to minimize technical artefacts and reduce the risk of false-positive findings. In specific clinical or research settings, additional laboratory validation may be performed for variants of particular interest. For example, in multiple sclerosis studies, when disease-associated genotypes cannot be reliably imputed from GWAS data, PCR-based assays targeting HLA-A2 and HLA-DR15 alleles may be carried out to confirm the presence of susceptibility-associated variants.

- Hardy-Weinberg equilibrium (HWE)

The Hardy-Weinberg equilibrium is a foundation principle in population genetics. It states that that in an ideal, undisturbed population, genetic variation will remain constant from one generation to the next. WHE is used as a baseline control, comparing real population's genetics against this perfect baseline, in order to identify genetic mutations or technical errors in data. SNPs showing statistically significant deviation from HWE, particularly within the control population, are excluded from the analysis[67].

- Minor allele frequency (MAF)

Minor allele frequency refers to the frequency at which the less common allele occurs at a specific location in the DNA within a given population. Variants with extremely low MAF are often removed, because rare alleles are more difficult to analyse reliably and are more susceptible to genotyping errors. Furthermore, low-frequency variants generally provide reduce statistical power in standard GWAS analyses, especially in studies with limited sample size[68].

6.2.3. Statistical association analysis

Once high-quality raw datasets have been obtained following QC procedures, the subsequent step in GWAS analysis is the identification of statistically significant SNPs associated with the phenotype of interest. In GWAS, association testing aims to evaluate whether the frequency of a specific allele or genotype differs significantly between affected individuals and unaffected controls. Since complex diseases are generally influenced by a large number of variants with small effect sizes, GWAS require robust statistical models capable of detecting subtle genetic effects while minimizing false-positive associations[69].

The statistical significance of each association is evaluated through *p*-values, which represent the probability of observing the detected association under the null hypothesis of no genetic effect; the smaller the *p*-value, the stronger the statistical association between the SNP and the phenotype. However, a GWAS simultaneously analyses hundreds of thousands to millions of variants across the genome, generating an enormous multiple-testing burden. Without appropriate statistical correction, a substantial number of associations would appear significant purely by chance. To control the false-positive rate, GWAS commonly apply a genome-wide significance threshold based on the Bonferroni correction. The conventional threshold adopted in human GWAS is: $p < 5 * 10^{-8}$. This stringent significance level substantially reduces the probability of spurious association arising from the large number of independent statistical test performed across the genome[70,71].

The overall distribution of GWAS association statistics is commonly evaluated using quantile-quantile (QQ) plots. These plots compare the observed distribution of *p*-values against the theoretical distribution expected under the null hypothesis of no association, which is typically represented by a diagonal reference line. Strong deviations in the upper-right tail indicate the presence of true disease-associated loci characterized by highly significant *p*-values. QQ plots are an essential tool for assessing the overall the validity of a GWAS and detecting potential systemic biases[72].

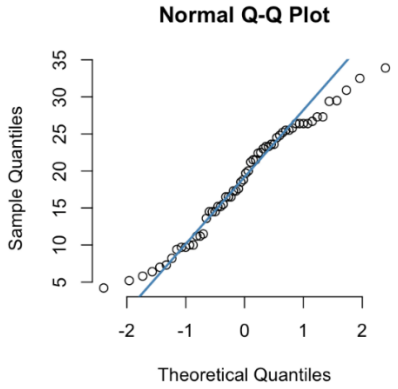

Figure 4. Quantile-Quantile (Q-Q) plot of observed versus expected *p*-values from the GWAS. The plot evaluates systematic bias and genomic inflation by comparing the observed association (*y*-axis) against distribution expected under the null hypothesis (*x*-axis). The close alignment of the data points along the diagonal reference line ($y = x$) indicates that the test statistics conform well to the null expectation, suggesting effective control for population stratification and confounding variables. No sharp upward deviation is observed at the right tail, indicating an absence of strong, genome-wide significant genetic associations in this analysis[68].

GWAS results are commonly visualized using a Manhattan plot, in which SNP association *p*-values are displayed on a $-\log_{10}$ scale along the *y*-axis, while the genomic position of SNPs across chromosomes are represented on the *x*-axis. The use of the $-\log_{10}$ transformation emphasizes very small *p*-values, allowing SNPs with the strongest statistical associations, and therefore potential disease-related variants, to be more easily identified[73].

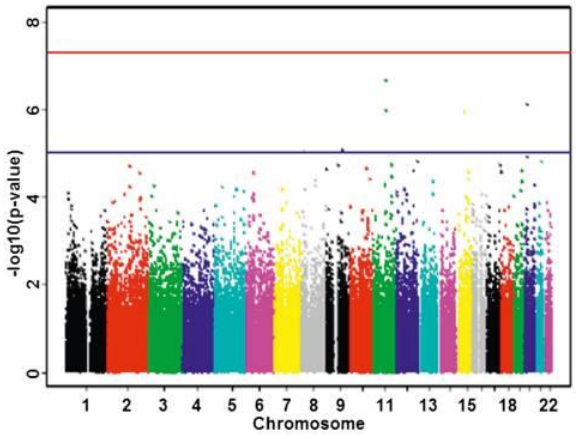

Figure 5. Manhattan plot of genome-wide association test statistics. The plot displays the genomic coordinates across autosomal chromosomes 1 to 22 (*x*-axis) plotted against the $-log_{10}$(*p*-value) of the association tests (*y*-axis). The horizontal blue line indicates the threshold for suggestive significance, while the upper horizontal red line denotes the threshold for strict genome-wide significance[69].

6.2.4. Polygenic Risk Score (PRS)

The identification of thousands of common genetic variants associated with complex diseases through GWAS has led to the development of polygenic risk scores (PRS), statistical models designed to estimate an individual's genetic susceptibility to a specific trait or disease. PRS are based on the assumption that complex disorders are highly polygenic, meaning that disease risk results from the cumulative contribution of numerous variants, each exerting only a modest effect on the phenotype. Rather than focusing only on genome-wide significant loci, PRS approaches aggregate the effects of a large number of variants distributed across the genome, including markers that individually may not reach statistical significance[74].

In PRS, genetic variants are first identified and subsequently combined into a single quantitative score. The score is generally calculated as the weighted sum of risk alleles carried by an individual, where each variant is weighted according to its estimated effect size derived from GWAS statistical analysis. This strategy allows the combined effect of multiple small-effect loci to be captured, thereby improving the estimation of inherited disease risk compared with single-variant analyses[75].

PRS do not function as deterministic predictors of disease onset, but rather as estimates of an individual's inherited genetic liability to a specific phenotype, meaning that environmental exposures and gene-environment interactions still contribute substantially to overall risk. However, as GWAS sample sizes expand, the accuracy of PRS is anticipated to improve. This advancement will refine disease stratification and progressively contribute to precision

medicine approaches, particularly by enabling the identification of high-risk individuals well before clinical symptoms manifest[76].

6.3. Limitations of GWAS

While GWAS have uncovered an unprecedent number of genetic variants linked to common traits and diseases, these markers explain only a modest fraction of the estimated heritability for most complex phenotypes. To account for this phenomenon, commonly referred to as missing heritability, several underlying mechanisms have been proposed. A major contributing factor is likely the exclusion of variants with subtle effect size that fail to cross the threshold required for genome-wide significance. Moreover gene-environment interaction are expected to explain at least part of the missing heritability, therefore, GWAS results should be interpreted in the context of environmental risk factors[77].

Another limitation of GWAS is the correlation of variants due to linkage disequilibrium. Although LD facilitates the initial identification of disease-associated loci, it can make it difficult to pinpoint the true causal variant or variants within the associated region. In addition, the use of SNP arrays introduces further limitations. Most arrays, particularly those used in early studies, were designed primarily using data from European populations, resulting in inadequate representation of linkage disequilibrium patterns in other ethnic groups and consequently poorer genomic coverage in non-European populations. More generally, because SNP arrays are based on predefined variants, they provide limited flexibility to expand genomic coverage and may therefore fail to detect novel SNP associations[77,78].

6.4. Contribution of GWAS to multiple sclerosis genetics

GWAS have become fundamental tools for investigating the genetic architecture of MS, enabling the identification of genetic variants associated not only with disease susceptibility, but also with clinical heterogeneity and therapeutic response. One of the earliest large-scale studies identified at least 48 new susceptibility loci, significantly expanding the number of known genetic risk factors for the disease[47]. Most of the genes located within these loci are involved in immune-related pathways, further supporting the central role of immune dysregulation in MS pathogenesis. Many MS-associated genes encode proteins participating in antigen presentation, T-cell activation, cytokine signalling, and other key mechanisms of adaptive and innate immunity[79].

The growing amount of GWAS data generated for MS and other complex disorders is systematically collected in publicly accessible databases such as the NHGRI-EBI GWAS Catalog, which continuously curates published genome-wide association studies and SNP-trait associations identified across human diseases and phenotypes. Large-scale international collaborations have played a crucial role in these discoveries by enabling the recruitment of sufficiently powered cohorts compose of thousands of patients and controls from multiple populations worldwide. Among the most important collaborative initiatives in MS genetics are the International Multiple Sclerosis Genetic Consortium (IMSGC), the Welcome Trust Case Control Consortium 2 (WTCCC2), and Australia and New Zealand Multiple Sclerosis Genetics Consortium, all of which have substantially contributed to the identification and validation of MS-associated genetic loci[79].

In 2007, the first GWAS conducted in MS patients was published, identifying the first genome-wide association outside the MHC region, located in the ILR2A gene. Several additional GWAS subsequently followed, with increased statistical power, larger cohorts, and, frequently, meta-analyses of case-control studies genotyped using different arrays[46,80]. To date, association studies have identified more than 200 susceptibility loci linked to MS risk. In particular, the largest study published by the International Multiple Sclerosis Genetics Consortium (IMSGC) in 2019 identified 233 statistically independent associations with MS susceptibility. Among these, 200 were autosomal susceptibility variants located outside the MHC region, 32 were located within the MHC region, and one variant was identified on the X chromosome[52].

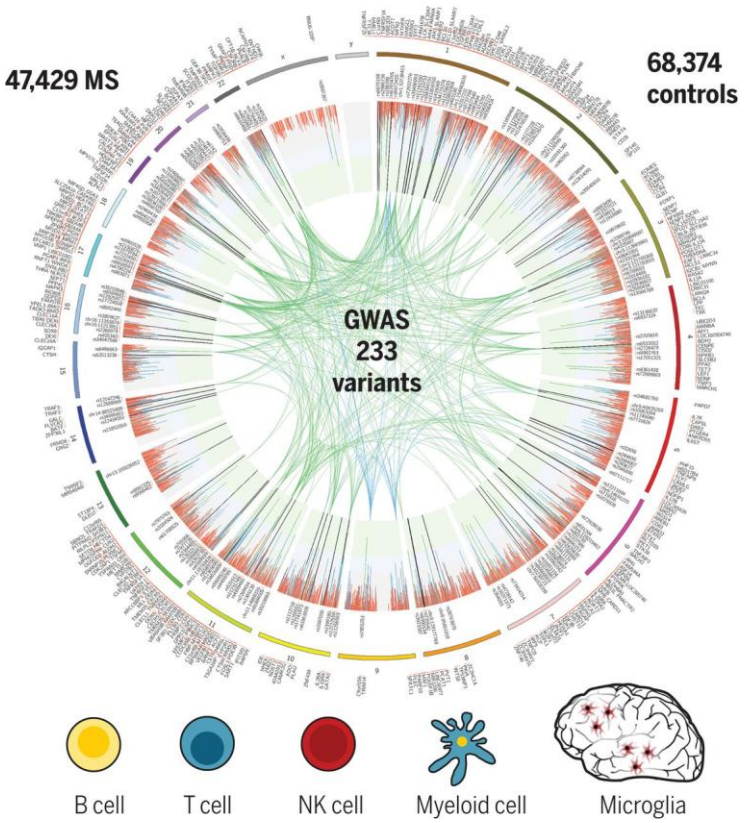

Figure 6. Circos plot representing the global genomic landscape of 233 multiple sclerosis (MS) risk variants. This circular visualization integrates the results of a large-scale GWAS comprising 47,429 MS cases and 68,374 controls. The outer ring displays autosomal chromosomes (1–22) and sex chromosomes (X, Y) labelled with specific associated single-nucleotide polymorphisms (SNPs) and mapping to candidate susceptibility genes. The bottom legend details color-coded functional annotations tying specific genetic variants to key target cells: B cells, T cells, NK cells, myeloid cells, and microglia, highlighting the heavily immune-mediated pathogenesis of the disease[80].

Overall, GWAS have profoundly advanced the understanding of the genetic basis of MS, revealing the highly polygenic nature of the disease and highlighting the predominant contribution of immune-related pathways to its pathogenesis. The continuous increase in sample size, together with international collaborative efforts and improved analytical approaches, has progressively expanded the number of identified susceptibility loci. These discoveries not only provide insights into the biological mechanisms underlying MS but also establish a foundation for the development of more precise diagnostic tools, personalized therapeutic strategies, and future functional studies aimed at clarifying the role of disease-associated variants.

Aim Of the Study

While genome-wide association studies (GWAS) have successfully mapped over 230 genetic variants to multiple sclerosis (MS) susceptibility, predicting the highly heterogeneous clinical course and long-term disability accumulation of the disease remains a challenge in clinical practice. Traditional genetic markers, such as the major risk allele HLA-DRB1*15:01, serve as powerful indicators for disease onset but show limited capacity to forecast individual trajectories of physical and cognitive deterioration. Furthermore, disease progression is increasingly understood to be a multifactorial process driven by an intricate interplay between an individual's genetic architecture and environmental modulators. Among these, cigarette smoking represents one of the most prominent modifiable lifestyle risk factors linked to both systemic immune dysregulation and chronic neuroaxonal injury. The aim of this study is to investigate whether individual genetic variables (including protective factors like HLA-A*02:01, progression loci like rs10191329, and neurodegenerative variants like APOE ε4), prominent environmental lifestyle factors (specifically cigarette smoking habit), and cumulative genomic risk scores can predict patterns of long-term disability accumulation in a large cohort of approximately 1,500 MS patients.

Specifically, this study evaluates the individual and collective capability of these variables to influence the accumulation of disability in patients, particularly through the analysis of relapse-associated worsening (RAW), progression independent of relapse activity (PIRA), and confirmed disability accumulation (CDA), thereby assessing the utility of these markers as predictive instruments to foster personalized clinical management and therapeutic strategies.

Clinical data regarding PIRA (Progression Independent of Relapse Activity), RAW (Relapse-Associated Worsening), and CDA (Confirmed Disability Accumulation) were available through a pilot linkage project between the PROGEMUS Consortium (PROgnostic GEnetic study in MUltiple Sclerosis), coordinated and managed by Professor Sandra D'Alfonso and Professor Maurizio Leone, and the Italian Multiple Sclerosis Registry (RISM). We would like to acknowledge and thank Professor Maria Trojano and Dr. Giuseppe Lucisano (University of Bari) for their contribution to this initiative. The project was designed to test and evaluate the feasibility of a procedure for linking the two databases, which contain extensive genetic and phenotypic data on individuals with multiple sclerosis.

Materials And Methods

1. EnviMS-Q

The EnviMS questionnaire was administered to all patients. This instrument collects information on environmental exposures that have been implicated in the etiology of multiple sclerosis. Although these factors are considered relevant across the lifespan, the questionnaire specifically investigates exposures occurring during childhood and adolescence (13–19 years of age). In particular, the questionnaire assessed dietary habits, sun exposure, medical history, active and passive smoking, body mass index (BMI), and hormonal factors.

The questionnaires had been scanned, and the information have been collected in a summary database. Of all the data available, we extracted information about smoke habit.

2. DNA extraction

DNA for GWAS analysis was extracted from blood samples collected from MS patients. Two different extraction methods were used: the Miniprep method and the Salting-Out technique. The MINI method was performed using the ReliaPrep™ Blood gDNA Miniprep System (Promega).

The MINI method allows rapid DNA extraction, although it generally yields lower-quality DNA compared with the Salting-Out protocol. In contrast, the Salting-Out technique is more time-consuming, requiring larger sample volume and a two-day procedure; however, it provides higher-quality DNA.

Since DNA quality and purity are critical parameters for GWAS, the extraction method was selected according to sample availability and the quality requirements of subsequent analyses.

2.1. Salting Out

Genomic DNA was extracted from peripheral blood samples using the Salting-Out protocol starting from 10 mL peripheral blood samples. Briefly, blood samples were thawed at room temperature. To perform the erythrocyte lysis, a haemolysis buffer containing $MgCl_2$, purified water, and NP-40 (Igepal) was prepared. In addition, a 0.9% saline solution was prepared.

Whole blood samples were transferred into 50 mL Falcon tubes and washed twice with 0.9% saline solution, followed by centrifugation at 2500 rpm for 30 minutes and subsequently for 20 minutes at room temperature. After each centrifugation step, the supernatant was carefully

discarded while preserving the cellular pellet. The pellet was then resuspended in haemolysis buffer to a final volume of 30 mL and incubated for 30 minutes at room temperature to promote erythrocyte lysis. Samples were centrifuged at 2500 rpm for 20 minutes, and the supernatant was removed. An additional wash with 0.9% saline solution was sequentially performed, followed by centrifugation at 1500 rpm for 10 minutes.

For cellular digestion, the resulting pellet was treated with Lysis Buffer (0.25 mL for each ml of starting blood volume), SDS (25 µL for each mL of Lysis Buffer), and proteinase K (25 µL for each mL of Lysis Buffer). Samples were incubated overnight at 37°C while shaking.

The following day, protein precipitation was performed using 6 M NaCl (270 µL for each mL of Lysis Buffer). Samples were vigorously shaken and centrifuged at 2500 rpm for 20 minutes at room temperature. After centrifugation, the supernatant was transferred into new Falcon tubes, avoiding the foam and leaving behind the pellet containing denatured proteins.

To precipitate the DNA, cold 100% ethanol (-20°C) was added to the supernatant for 2 minutes, until visible DNA filaments formed. Using a glass hook, the DNA was recovered and subsequently washed three times with 70% ethanol. The hook was then placed upside-down to allow complete ethanol evaporation from the DNA sample, generally achieved within 2-3 hours.

After complete ethanol evaporation, the DNA was transferred into new Eppendorf tubes containing TE buffer. The volume of TE buffer was estimated according to the size of the DNA filaments and generally ranged from 50 to 400 µL. DNA was left in TE buffer for approximately 2-3 hours to allow complete solubilization. The Eppendorf tubes containing DNA and TE buffer were then kept shaking for 20-30 minutes and stored overnight at 4°C.

DNA concentration and purity were assessed using a NanoDrop spectrophotometer.

2.2. Miniprep System

Genomic DNA was extracted from peripheral blood samples using the ReliaPrep™ Blood gDNA Miniprep System (Promega) from 200 µL aliquots of peripheral blood.

Proteinase K (20 µL) was added to each aliquot to digest proteins, followed by the addition of 200 µL of lysis buffer to induce membrane disruption. Samples were briefly vortexed, centrifuged, and subsequently incubated in a water bath at 56°C for 10 minutes.

Binding Buffer (250 µL) was then added, and samples were vortexed for at least 10 seconds. Samples were subsequently transferred into membrane spin columns placed in collection tubes and centrifuged at 13,000 rpm for 1 minute.

The collection tubes were discarded, and the membrane spin columns were transferred into new tubes. Two subsequent washing steps were performed using 500 µL of washing buffer, followed by centrifugation at 13,000 rpm for 3 minutes.

An additional centrifugation step with empty columns was performed at 13,000 rpm for 1 minute. The membrane spin columns were then transferred into new Eppendorf tubes, and 50 µL of nuclease-free water was added.

A final centrifugation step was performed at 13,000 rpm for 1 minute. The membrane spin columns were discarded, and the DNA samples were stored in the Eppendorf tubes. DNA concentration and purity were assessed using a NanoDrop spectrophotometer.

3. Purification and recovery of DNA

This procedure was performed on DNA samples containing salts or organic contaminants, with a minimum starting volume of 300 µL TE buffer.

All procedures involving phenol and chloroform were performed under a chemical hood. Equal volumes of phenol and chloroform (pH 8.0) were added to the DNA samples, each corresponding to half of the initial DNA volume. After vortexing, the samples were centrifuged at 13,000 rpm for 15 minutes.

The supernatant was carefully collected, and a volume of chloroform equal to the recovered supernatant volume was added to the Eppendorf tubes. Samples were then vortexed and centrifuged at 13,000 rpm for 10 minutes.

After collection of the supernatant, two volumes of cold 100% ethanol and 0.1 volume of 3 M sodium acetate (pH 5.2) were added to the samples. Samples were incubated at -80°C for 20 minutes and subsequently centrifuged at 13,000 rpm for 15 minutes.

The supernatant was removed by aspiration, and the pellet was washed with cold 70% ethanol. Samples were then centrifuged at 13,000 rpm for 10 minutes.

After removal of the supernatant, the pellet was allowed to air-dry for 24 hours and was subsequently resuspended in TE buffer.

4. DNA amplification through Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR) was performed using genomic DNA as template to amplify the target gene regions. Genomic DNA was previously extracted from peripheral blood samples using the Salting-Out and the MINI methods and quantified using a NanoDrop spectrophotometer.

The PCR mix contains all the reagents required to selectively amplify a specific DNA target region. Each reagent plays an essential role in ensuring the efficiency, specificity and reliability of the amplification process.

The reaction buffer is necessary to maintain optimal chemical conditions throughout the PCR reaction, essential for proper enzymatic activity. The dNTPs (deoxynucleotide triphosphates) represent the building blocks used by the DNA polymerase to synthesize new DNA strands during amplification. Magnesium chloride ($MgCl_2$) acts as a cofactor for Taq DNA polymerase and is required for enzymatic activity. In addition, magnesium concentration strongly influences PCR specificity and amplification efficiency.

The primers are designed to anneal to complementary regions flanking the target DNA sequence. These primers determine the specificity of the reaction by defining the DNA region to be amplified. Taq DNA polymerase is responsible for synthesizing new DNA strands through the incorporation of complementary nucleotides during the extension phase. Finally, genomic DNA extracted from peripheral blood samples serves as the template containing the target sequence to be amplified.

PCR analyses were performed on samples that failed GWAS quality control due to sex discrepancies between reported and genetically inferred sex (Amelogenin) or excessive SNP missingness (HLA-DR15/15*01 and HLA-A2).

4.1. Amelogenin PCR assay for sex determination

This assay allows to determine the chromosomal sex of a sample, and it can be used as a quality step prior to the genotyping phase or to resolve sex-mismatches between the declared sex and the sex determined in a GWAS experiment (basing on the heterozygous state of SNPs located on X chromosome), that may be due to errors in sample handling.

The assays is a multiplex PCR that amplifies two distinct fragments:

- SRY single copy gene present only on chromosome Y, that allows the determination of the sex
- AMEL (amelogenin gene) is located on the pseudo-autosomic portion of the X and Y chromosomes, but the copy located on chromosome Y (AMELY) is 6 pb longer compared to AMELX.

A male sample is going to present two band, corresponding to SRY (197 bb) and AMELY (112 bp), while a female sample is going to present only the AMELX band (106 bp).

PCR reaction was prepared in a final volume of 15 µL containing 13 µL of PCR Mix (**Table 1**) and 2 µL of genomic DNA (50 ng). The Primer Mix was prepared using the volumes shown in **Table 3**.

Reagents	Initial concentrations	Final concentrations	Reaction Mix
GoTaq® G2 Flexi Buffer	5X	1X	3 µl
dNTPs	2,5 mM	0,2 mM	1,2 µl
$Mgcl_2$	25 mM	1,5 mM	0,9 µl
Primer Mix AMEL + SRY	10 pmol/µl	0,48 pmol/µl	0,72 µl
GoTaq® Polymerase (Promega)	5 U/µl	0,024 U/µl	0,072 µl
H_2O	-	-	7,108 µl
Final Volume			13 µl
+ 2 µl DNA [25 ng/ µl]			

Table 1. Reaction components and formulation of the Amelogenin PCR master mix. The table outlines the initial and final concentrations of each reagent, along with the specific volumes required for a single amplification reaction. The volumes are calculated based on a final reaction volume of **15** µL.

Amplification was carried out in a thermal cycler under the following conditions: an initial denaturation step at 94°C for 5 minutes, followed by 32 cycles of denaturation at 94°C for 20 seconds, primer annealing at 65°C for 30 seconds, and extension at 72°C for 20 seconds. A final extension step was performed at 72°C for 7 minutes (**Table 2**).

PCR products were analysed by electrophoresis on a 2% agarose gel containing SYBR Safe and visualized using a ChemiDoc imaging system.

THERMAL CYCLING PROTOCOL		
Temperature	**Time**	**Cycles**
94°C	5’	1
94°C	20”	32
65°C	30”	
72°C	20”	
72°C	7’	1
4°C	∞	

Table 2. Thermal cycling protocol for Amelogenin PCR reaction

Primers	**Initial concentrations**	**Volumes**
AMEL [F]	10 pmol/µl	10 µl
AMEL [R]	10 pmol/µl	10 µl
SRY [F]	10 pmol/µl	10 µl
SRY [R]	10 pmol/µl	10 µl
Total		40 µl
+ 60 µl of water		

Table 3. Primer Mix components.

4.2.HLA-DRB1*15 assay

This is a multiplex-allele specific PCR assay that uses two couples of primers:

1) DR15 primers: specifically designed to bind only to the DR15 allele. The corresponding amplicon is amplified only in patients that are DR15 positive (DR15+/DR15+ or DR15+/DR15-) and it is 196 bp long.

2) MOG primers: this amplicon (475pb) is used as internal control.
   This assay allows to determine if a sample is positive or negative for DR15, but not to determine if it is heterozygous or homozygous.

PCR reaction was prepared in a final volume of 15 µL containing 13 µL of PCR Mix (**Table 4**) and 2 µL of genomic DNA (50 ng).

Reagents	Initial concentrations	Final concentrations	Reaction Mix
GoTaq® G2 Flexi Buffer	5X	1X	3 µl
dNTPs	2,5 mM	0,2 mM	1,2 µl
$Mgcl_2$	25 mM	1,5 mM	0,9 µl
Oligo Mix MOG [5]	5 pmol/µl	0,25 pmol/µl	0,75 µl
Oligo Mix DR15 [10]	10 pmol/µl	0,5 pmol/µl	0,75 µl
GoTaq® Polymerase (Promega)	5 U/µl	0,02 U/µl	0,06 µl
H_2O	-	-	6,34 µl
Final Volume			13 µl
+ 2 µl DNA [25 ng/ µl]			

Table 4. Reaction components and formulation of the HLA-DR15 PCR master mix. The table outlines the initial and final concentrations of each reagent, along with the specific volumes required for a single amplification reaction. The volumes are calculated based on a final reaction volume of 15 µL.

Amplification was carried out in a thermal cycler under the following conditions: an initial denaturation step at 96°C for 5 minutes, followed by 30 cycles of denaturation at 96°C for 30 seconds, primer annealing at 60°C for 1 minute, and extension at 72°C for 1 minute and 30 seconds. A final extension step was performed at 72°C for 5 minutes (**Table 5**).

PCR products were analysed by electrophoresis on a 1.5% agarose gel containing SYBR Safe and visualized using a ChemiDoc imaging system.

THERMAL CYCLING PROTOCOL		
Temperature	**Time**	**Cycles**
96°C	5’	1
96°C	30”	30
60°C	1’	
72°C	1’30”	
72°C	5’	1
10°C	∞	

Table 5. Thermal cycling protocol for HLA-DR15 PCR reaction

4.2.1. HLA-DRB1*15:01

Samples that were positive for the HLA-DRB1*15 assay have been further characterized to discriminate between the two most common alleles: DRB1*15:01 or DRB1:15:02 using a multiplex allele-specific PCR assay specific for DRB1*15:01 allele.

PCR reaction was prepared in a final volume of 15 µL containing 13 µL of PCR Mix (**Table 6**) and 2 µL of genomic DNA (50 ng).

Amplification was carried out in a thermal cycler under the following conditions: an initial denaturation step at 96°C for 5 minutes, followed by 30 cycles of denaturation at 96°C for 30 seconds, primer annealing at 60°C for 1 minute, and extension at 72°C for 1 minute and 30 seconds. A final extension step was performed at 72°C for 5 minutes (**Table 7**).

PCR products were analysed by electrophoresis on a 1.5% agarose gel containing SYBR Safe and visualized using a ChemiDoc imaging system.

Reagents	Initial concentrations	Final concentrations	Reaction Mix
GoTaq® G2 Flexi Buffer	5X	1X	3 µl
dNTPs	2,5 mM	0,2 mM	1,2 µl
$Mgcl_2$	25 mM	1,5 mM	0,9 µl
Oligo Mix MOG [5]	5 pmol/µl	0,25 pmol/µl	0,75 µl
Oligo Mix DR15*01 [10]	10 pmol/µl	0,5 pmol/µl	0,75 µl
GoTaq® Polymerase (Promega)	5 U/µl	0,02 U/µl	0,06 µl
H_2O	-	-	6,34 µl
Final Volume			13 µl
+ 2 µl DNA [25 ng/ µl]			

Table 6. Reaction components and formulation of the HLA-DR15*01 PCR master mix. The table outlines the initial and final concentrations of each reagent, along with the specific volumes required for a single amplification reaction. The volumes are calculated based on a final reaction volume of 15 µL.

THERMAL CYCLING PROTOCOL		
Temperature	**Time**	**Cycles**
96°C	5’	1
96°C	30”	30
60°C	1’	
72°C	1’30”	
72°C	5’	1
10°C	∞	

Table 7. Thermal cycling protocol for HLA-DR15*01 PCR reaction

4.3. HLA-A*02 assay

This assay includes a PCR that amplifies the whole exon 2 of HLA-A gene, followed by an enzyme digestion. PCR reaction was prepared in a final volume of 15 µL containing 13 µL of PCR Mix (**Table 8**) and 2 µL of genomic DNA (100 ng).

Amplification was carried out in a thermal cycler under the following conditions: an initial denaturation step at 96°C for 5 minutes, followed by 35 cycles of denaturation at 96°C for 30 seconds, primer annealing at 62°C for 30 seconds, and extension at 72°C for 30 seconds. A final extension step was performed at 72°C for 5 minutes (**Table 9**).

PCR products were analysed by electrophoresis on a 1.5% agarose gel containing SYBR Safe and visualized using a ChemiDoc imaging system.

Reagents	Initial concentrations	Final concentrations	Reaction Mix
GoTaq® G2 Flexi Buffer	5X	1X	3 µl
dNTPs	2,5 mM	0,2 mM	1,2 µl
$Mgcl_2$	25 mM	1,75 mM	1,05 µl
Primer 5' deg F [5]	5 pmol/µl	0,2 pmol/µl	0,6 µl
Primer IntEx 2 [5]	5 pmol/µl	0,2 pmol/µl	0,6 µl
GoTaq® Polymerase (Promega)CONTROLLARE	10 U/µl	0,05 U/µl	0,08 µl
H_2O	-	-	6,47 µl
Final Volume			13 µl
+ 2 µl DNA [25 ng/ µl]			

Table 8. Reaction components and formulation of the HLA-A*02 PCR master mix. The table outlines the initial and final concentrations of each reagent, along with the specific volumes required for a single amplification reaction. The volumes are calculated based on a final reaction volume of 15 µL.

THERMAL CYCLING PROTOCOL		
Temperature	**Time**	**Cycles**
96°C	5’	1
96°C	30”	35
62°C	30”	
72°C	30”	
72°C	5’	1
10°C	∞	

Table 9. Thermal cycling protocol for HLA-A*02 PCR reaction

The PCR product was digested with the restriction enzyme BspEI, to distinguish between HLA-A*02 alleles. The enzyme cuts a 5’-TCCGGA-3’ sequence present only on HLA-A:02 allele. The amplicon (215 bp) is cut in two fragments (160 bp and 55 bp).
The digestion reaction was prepared in a final volume of 20 µL containing 10 µL of Digestion Mix (**Table 10**) and 10 µL of PCR product.
The samples were incubated in a thermal cycler at 37°C for 4 hours to allow restriction enzyme activity, and BspEI was then inactivated at 80°C for 20 minutes.

Samples were analysed by electrophoresis on a 2,5% agarose gel containing SYBR Safe and visualized using a ChemiDoc imaging system.

Primers	**Initial concentrations**	**Volumes for 1 sample**
Cut Smart Buffer (NEB)	10X	2 µl
BspEI enzyme	10.000 U/ml	1 µl
H_2O		7 µl
+ 10 µl of PCR product		

Table 10. Digestion Mix components.

5. DNA sequencing

PCR products were purified enzymatically using ExoSAP method, in order to remove residual primers and unincorporated dNTPs prior sequencing analysis. Exonuclease I was used to degrade excess single-stranded primers, while Shrimp Alkaline Phosphatase (SAP) dephosphorylated residual dNTPs.

For each reaction, a purification mix containing nuclease-free water, Exonuclease I, and SAP was prepared. Approximately 1 µL of PCR product was added to 5 µL of purification mix. Samples were incubated at 37°C for 15 minute to activate the enzymatic digestion, followed by incubation at 80°C for 15 minutes to inactivate the enzymes.

Purified PCR products were subjected to Sanger sequencing using the BigDye Terminator chemistry. Sequencing reactions were prepared using BigDye reagent, sequencing buffer, purified PCR product, and either forward or reverse primers at a final concentration of 3.2 µM.

Each sequencing reaction had a final volume of 10 µL, consisting of sequencing mix (6 µL), primer solution (1 µL), and purified PCR product (3 µL). Amplification was performed in a thermal cycler according to thermal protocol in **Table 11**. The sequencing reaction generated fluorescently labelled DNA fragments of different lengths, allowing sequence determination by capillary electrophoresis.

Following cycle sequencing, DNA products were purified by ethanol precipitation to remove salts and unincorporated fluorescent dyes. 1 µL of sodium acetate and 30 µL of cold 100% ethanol were added to each sample and then left at -20°C for 20 minutes. Samples were centrifuged at 4.000 rpm for 30 minutes at 4°C, subsequently the supernatant was discarded.

Two subsequent steps of purifications with 30 µL of cold 70% ethanol were performed, and samples were centrifuged at 4.000 rpm for 15 minutes at 4°C. After the last centrifuging step, supernatant was carefully removed, and samples were left at 96°C for approximately 5 minutes, to evaporate residual ethanol. Purified DNA pellets were resuspended in 10 µL of formalmide to maintain single-stranded DNA linearity.

For sequencing analysis, 10 µL of formalmide were dispensed into each well of the sequencing plate, followed by the addition of purified sequencing products. Samples were denatured at 96°C for 2 minutes and immediately cooled on ice before loading onto the capillary sequencer.

Sequencing data were analysed using the program Chromas. Electropherograms were evaluated based on fluorescence peak intensity to confirm sequence quality and identify nucleotides variations.

THERMAL CYCLING PROTOCOL		
Temperature	**Time**	**Cycles**
96°C	3’	1
96°C	20”	25
50°C	5”	
60°C	4’	
4°C	∞	

Table 11. Thermal cycling protocol for sequencing reaction

6. Calculation of Polygenic Risk Scores (PRS)

Genetic and clinical data from approximately 1,500 patients were obtained from a genome-wide association study (GWAS) dataset. Genetic variables included the presence or absence of HLA-DRB1*15:1019132901, HLA-A*02:01, APOE ε4, and the severity-associated SNP rs10191329. APOE alleles have been reconstructed basing on the genotypes of the 2 SNP

In addition, 7 PRS scores were calculated:

A. 6 “susceptibility scores”, including SNPs already identified as loci statistically associated with MS, according to the GWAS of 2019 performed by the International Multiple Sclerosis Genetics Consortium (IMSGC)[52]; in this study 200 non-HLA loci with statistical association with MS susceptibility ($P < 5 \times 10^{-8}$) and the 32 HLA markers have been identified. Moreover, loci which did not reached the standard threshold of GWAS significance, but with available replication data, were identified by the IMSGC and classified as strongly suggestive effects (sS; n=117) and underpowered suggestive effects (unS; n=299). sS were used to calculate scores named as “Sigstrong”, whereas unS were used for scores named as “Sigstrongweak”.

Here the classification of the different scores, with reported the genetic information used to calculate them:

- Sig_score (200 NON-HLA loci)
- Sig_hla (200 NON-HLA loci + HLA markers)
- Sigstrong_score (200 NON-HLA loci + 117 sS SNP)
- Sigstrong_hla (200 NON-HLA loci + 117 sS + HLA markers)
- Sigstrongweak_score (200 NON-HLA loci + 117 sS + 299 unS)

Sigstrongweak_hla (200 NON-HLA loci + 117 sS + 299 unS + HLA markers)

The web-based application LDProxy (https://ldlink.nih.gov/ldproxy) was used to retrieve the relevant best proxies for the SNPs not present in our dataset. Specifically, we set EUR population and a window of 500,000 bp for the LD R2 calculation and retained only the proxies with R2 > 0.80. This reduced the total list to 198 SNPs for the Sig score, 300 SNPs for the SigS score, and 541 SNPs for the SigSW score.

All these scores have been generated by determining for each individual, at each genetic variant, the number of risk alleles (0, 1, or 2 in case of homozygosity for protective the allele, heterozygosity or homozygosity for the risk allele). This number has been multiplied by the "weight" of the SNP, determined as the ln (OR) conferred by the SNP (IMSGC). Finally, the PRS has been calculated as an arithmetic sum on all the loci included in the score. The calculation is performed with a custom script written in R v4.5.1.

B. A "severity score" including the rs10191329 plus 9 other variants associated at suggestive level with disease severity, basing on a recent GWAS (IMSGC 2023[81]) that analysed association with ARMSSS (Age-Related Multiple Sclerosis Severity Score), a clinical and prognostic score.

The score was calculated similarly as described for the 6 "susceptibility scores".

7. Statistical Analysis

The clinical outcomes analysed were relapse-associated worsening (RAW), progression independent of relapse activity (PIRA), and confirmed disability accumulation (CDA). Time-to-event variables were defined as the number of days from baseline to the occurrence of the corresponding confirmed disability progression event. Patients who did not experience the event during follow-up were considered censored at the date of the last available assessment.

Associations between genetic variables and time to-event outcomes were evaluated through univariate survival analyses and survival analyses covaried by sex and age at the disease onset performed with the statistical software R. Cox proportional hazards regression models were used to assess the relationship between each genetic marker and the risk of developing RAW, PIRA or CDA. For HLA-DR15, HLA-A2, APOE ε4, and the severity-associated SNP rs10191329, patients were stratified according to the presence or absence of the respective genetic variant.

Kaplan-Meier survival curves were generated and graphically represented to visualise differences in event-free survival between groups. For genetic score analyses, patients were categorized into quartiles according to the distribution of each score. Survival analyses were subsequently performed by comparing individuals in the first quartile (lowest genetic burden) with those in the fourth quartile (highest genetic burden). Separate analyses were conducted for RAW, PIRA, and CDA outcomes.

Furthermore, for HLA-DR15, HLA-A2, APOE ε4 we calculated association between presence of the risk allele and having reached the progression outcome constructing 2X2 contingency tables. The risk conferred by the genetic variant was determined as Odd Ratio (OR) with its corresponding 95% Confidence Interval (CI) and p-value.

In addition, genetic score distributions were compared between patients with and without each disability progression phenotype (PIRA+/PIRA-, RAW+/RAW-, and CDA+/CDA-). Mean values and standard deviations were calculated for each group, and differences between groups were assessed using independent two-sample t-test.

Information about smoke habit were dichotomized in one variable (ever been regular smoker/not smoker). Association between smoking status and clinical outcomes were performed as described for dichotomic genetic variants.

Results

1. General workflow of the study.

This study was aimed at testing genetic and environmental markers for their influence on the accumulation of disability in MS patients.

We used a set of 1552 MS patients (already collected by the PROGEMUS consortium, coordinated by prof. S. D'Alfonso and prof. M. Leone) genotyped at whole genome level and for which we have availability of detailed clinical information (thanks to the linkage with RISM, coordinated by prf. M Trojano), including times at PIRA, RAW and CDA.

The workflow of the study includes:

- Quality checks of the samples prior whole-genome genotyping
- Genome-wide genotyping: most samples had GWAS data previously available thanks to projects previously conducted in the genetic laboratory. During my thesis a sample set of 650 samples were genotyped at the Biodiversa institute (Verona), using the array Illumina GSA (global screening array) containing about 650K SNPs
- Post-Genotyping Quality Controls of the samples.
- Whole-genome genotype imputation and Imputation of classical HLA alleles have been performed by dr. Nicola Pomella. For all these samples, an imputation step has been performed on the "Haplotype Reference Consortium Rel 1.1 2016 Eurall(EUR) Panel", allowing inference for genotypes of ~8 million variants.
- Selection of Genetic and environmental variants for the analysis: the sample set is not powered for an unbiased association analysis on all the variants, therefore we decided to select a number of HLA and not-HLA genetic variants already reported for association with susceptibility to MS or with disease severity, and to construct Polygenic Risk Scores (PRS) that includes those variants. Namely, the tested variants are HLA-DRB1*15:1019132901, HLA-A*02:01, APOE ε4, and the severity-associated SNP rs10191329, plus 7 PRS.
- For three variants that are of particular interest (HLA-A*02, HLA-DRB1*01) we proceeded to laboratory genotyping of those samples that failed the GWAS+imputation pipeline

- Calculation of PRS.
- Statistical analysis

2. Quality check through amelogenin assay

As part of the quality checks performed prior the genotyping phase, a subset of 99 samples had been checked for DNA integrity and sex confirmation with an amelogenin PCR assay. The differential amplification of X (Amelogenin) and Y (SRY) chromosome-specific bands allowed for the unambiguous sex confirmation. By comparing the migration profiles against a 100 bp DNA ladder, the true biological sex of the samples was successfully validated, ensuring data integrity before downstream genomic analyses. The amelogenin assay confirmed the declared sex for all the samples (45 males and 54 females), therefore they have been included in the genomic analysis. Results are shown in **Figure 7**.

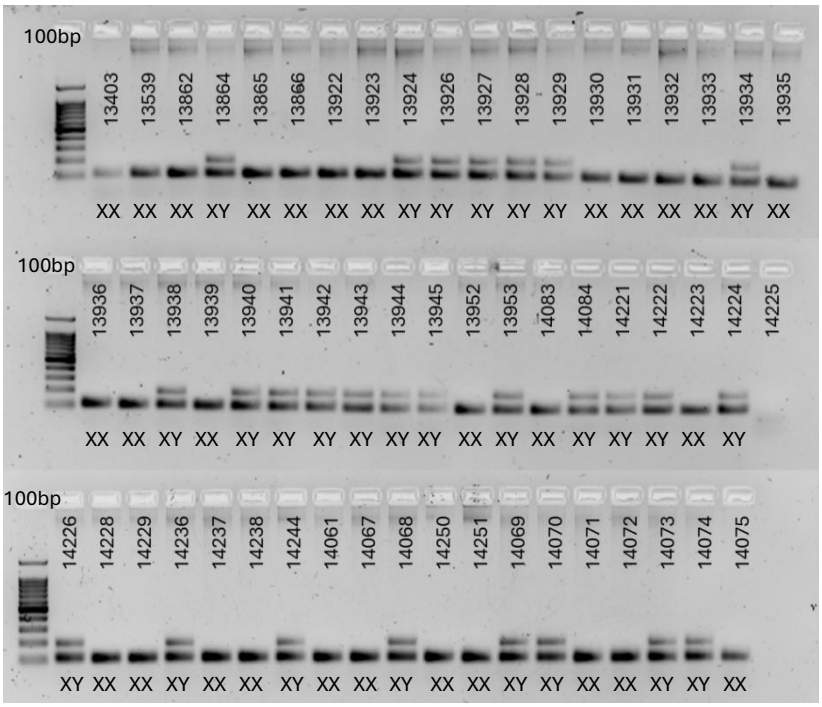

Figure 7. Amelogenin PCR assay for sex validation. PCR products resolved on a 2% agarose gel. The lower band (106 – 112 bp) correspond to AMEL gene and is present both in male and female samples. The band corresponding to AMELY is actually 6bp bigger than the one corresponding to AMELX, but this difference cannot be resolved with this gel. The upper band (200 bp) correspond to the SRY gene and is present only in male samples and allows to determine the sex.

3. HLA-DRB1*15:01 genotyping

A laboratory genotyping was conducted on 90 DNA samples that failed the imputation pipeline using two allele-specific PCR:

1) First, all samples have been analysed with the HLA-DRB1*15 assay. The successful amplification of the HLA-DR15 susceptibility allele was verified on a 1.5% agarose gel (**Figure 8**). This targeted experimental rescue allowed to recover crucial genotypic information that would have otherwise been excluded from the study. All samples resulting HLA-DR15 positive, where then analysed with a PCR specific for HLA-DRB1*15:01. 62 samples resulted HLA-DRB1*15 positive and 28 negative.

2) All positive samples have been subsequently genotyped with the HLA-DRB1*15:01 assay. (**Figure 9**) 56 samples resulted HLA-DRB1*15:01 positive.

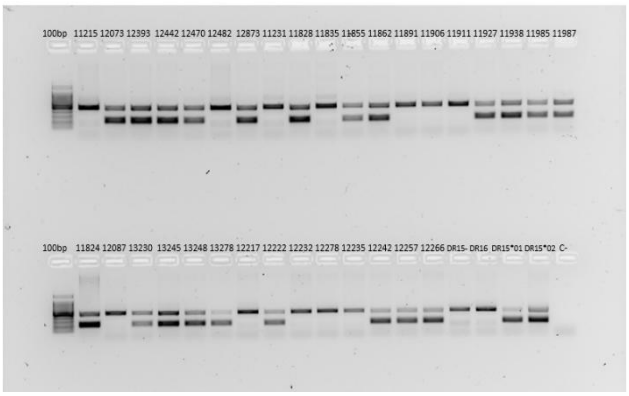

Figure 8. PCR amplification of the HLA-DR*15 susceptibility allele. Electrophoresis on a 1.5% agarose gel used as laboratory validation for missing GWAS data.

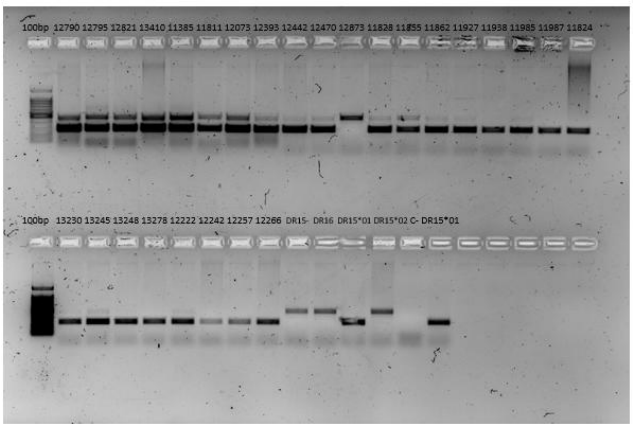

Figure 9. Verification of the HLA-DRB1*15:01 specific allele. 1.5% agarose gel electrophoresis confirming the presence of the genetic susceptibility variant. Samples presenting two bands are HLA-DRB1*15:01 positive.

4. HLA-A*02 PCR

A laboratory genotyping was conducted on 55 DNA samples that failed the imputation pipeline (**Figure 10**). 33 samples resulted heterozygote for HLA-A*02, 11 homozygotes and 11 negatives.

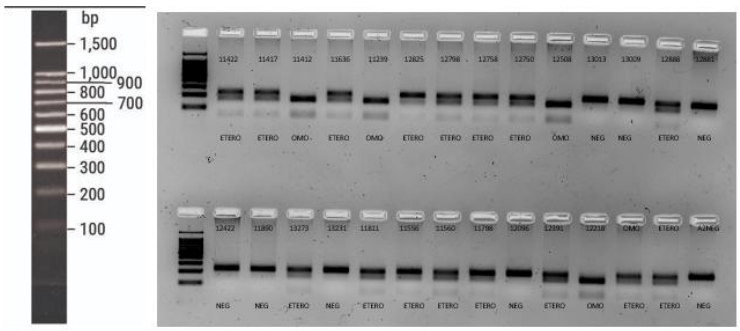

Figure 10. BspEI restriction profile for HLA-A*02 allele discrimination.
PCR products analysed on a 2.5% agarose gel following enzymatic digestion.

5. Association between genetic variants and clinical phenotypes

As first step, the association between genetic variants with clinical phenotypes was determined by comparing the presence/absence of the risk allele in MS patients that reached or not reached three disability progression outcomes (PIRA, RAW, and CDA). This was analysed with 2x2 contingency tables in a univariate model.

	Clinical phenotype					
Genetic variant	**RAW**		**PIRA**		**CDA**	
	OR (95% CI)	**P-value**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**
HLA-A2	0,75 (0,46-1,24)	0,2636	1,34 (1,05-1,73)	0,0215	1,23 (0,96-1,57)	0,1038
HLA-DR15	1,64 (1,05-2,56)	0,0292	1,16 (0,89-1,51)	0,2708	1,33 (1,03-1,72)	0,0263
APOE ε4	1,18 (0,66-2,12)	0,5681	1,14 (0,82-1,58)	0,4389	1,19 (0,86-1,63)	0,2949
rs10191329 (DYSF-ZNF638	1,70 (1,10-2,64)	0,0171	0,92 (0,71-1,19)	0,5063	1,08 (0,85-1,39)	0,5207

Table 12. Association between key genetic variants and clinical disability progression phenotypes. The table presents the results of a univariate statistical analysis evaluating the impact of genetic markers on Relapse-Associated Worsening (RAW), Progression Independent of Relapse Activity (PIRA), and Confirmed Disability Accumulation (CDA). Statistically significant associations ($P < 0.05$) are highlighted in yellow.

The comprehensive findings are summarized in **Table 12**.

The statistical evaluation revealed distinct, variant-specific associations with the different progression pathways, suggesting that separate genetic architectures may drive inflammatory relapse-mediated damage versus neurodegenerative, relapse-independent progression.

The HLA-A2 allele demonstrated a selective, statistically significant association with an increased risk of PIRA (OR = 1.34, 95% CI: 1.05–1.73, P = 0.0215). Notably, HLA-A2 did not demonstrate a significant correlation with RAW (P = 0.2636) or overall CDA (P = 0.1038). This selective correlation underscores a potential role for HLA-A2 in mechanisms underlying insidious neurodegeneration or compartmentalized central nervous system inflammation, which typically characterize progression independent of acute clinical relapses.

In contrast, the HLA-DR15 variant was significantly associated with heightened risks in both the RAW (OR = 1.64, 95% CI: 1.05-2.56, P = 0.0292) and CDA (OR = 1.33, 95% CI: 1.03–1.72, P = 0.0263) cohorts. Because CDA encompasses disability accumulation from both relapse-dependent and relapse-independent pathways, the strong signal observed in HLA-DR15 is likely heavily driven by its prominent effect on inflammatory relapse activity (RAW). This aligns with its well-established role in classical immune-mediated focal demyelination.

The Severity SNP exhibited the strongest effect size strictly within the relapse-mediated pathway, showing a significantly elevated risk for RAW (OR = 1.70, 95% CI: 1.10–2.64, P = 0.0171). No significant correlations were observed for this SNP regarding PIRA (P = 0.5063) or CDA (P = 0.5207), implying its pathophysiological impact may be confined to modulating the severity or poor recovery of acute relapse events rather than driving insidious progression.

Finally, the APOE ε4 variant did not reach statistical significance for any of the clinical phenotypes evaluated (RAW: P = 0.5681; PIRA: P = 0.4389; CDA: P = 0.2949). Within this specific cohort and univariate framework, APOE does not appear to act as a primary driver of short-to-medium-term disability progression.

6. Association between PRSs and clinical phenotypes

		Sig_score	Sig_hla	Sigstrong_score	Sigstrong_hla	Sigstrongweak_score	Sigstrongweak_hla	Severity score
RAW	Positive	90	90	90	90	90	90	90
	Negative	1110	1110	1110	1110	1110	1110	1118
	Mean positive	21,69	22,91	31,35	32,57	64,64	65,85	0,89
	Mean negative	21,82	23,01	31,45	32,65	64,59	65,78	0,87
	p-value	0,2384	0,4512	0,4141	0,5968	0,7642	0,6975	0,1787
PIRA	Positive	411	411	411	411	411	411	415
	Negative	789	789	789	789	789	789	793
	Mean positive	21,83	23,08	31,47	32,73	64,59	65,85	0,88
	Mean negative	21,80	22,96	31,43	32,59	64,60	65,76	0,87
	p-value	0,5875	0,0880	0,5233	0,0918	0,9412	0,3700	0,0430
CDA	Positive	501	501	501	501	501	501	505
	Negative	699	699	699	699	699	699	703
	Mean positive	21,80	23,05	31,45	32,70	64,60	65,85	0,88
	Mean negative	21,81	22,96	31,44	32,60	64,59	65,75	0,87
	p-value	0,9143	0,2154	0,8587	0,1804	0,9289	0,2844	0,0077

Table 13. Comparison of Polygenic Risk Scores (PRSs) across clinical phenotypes. The table displays patient counts, mean scores, and statistical significance levels (P-values) evaluating the association between different PRSs (including HLA and severity metrics) and disability progression outcomes: Relapse-Associated Worsening (RAW), Progression Independent of Relapse Activity (PIRA), and Confirmed Disability Accumulation (CDA). Statistically significant associations ($P < 0.05$) are highlighted in yellow. Positives: n. of patients that reach the outcome; negatives: n. of patients that did not reach the outcome; mean positives: mean in patients that reached the outcome (PIRA, RAW, or CDA); mean negatives: mean in patients that did not reach the outcome.

This section evaluates whether Polygenic Risk Scores (PRSs) and a specific Severity score can serve as predictors for the different pathways of disability progression (RAW, PIRA, and CDA). As first step we compared the mean PRS in patients that reached or not reached the three disability outcomes (PIRA; RAW; CDA) using a t-test analysis. The objective is to determine if the cumulative genetic burden or targeted severity metrics influence the pattern of clinical worsening. The detailed quantitative data, sample sizes, and corresponding significance levels (P-values) are presented in **Table 13.**

The Severity score is the only metric that demonstrates a nominally significant association with disability progression. Specifically, it is associated with overall disability accumulation (CDA, $P = 0.0077$) and progression independent of relapse activity (PIRA, $P = 0.0430$). Conversely, no significant association was found with relapse-associated worsening (RAW, $P = 0.1787$). This profile suggests that the Severity score acts as a selective indicator of the neurodegenerative mechanisms underlying PIRA, rather than the acute inflammatory processes typical of relapses.

None of the evaluated susceptibility risk scores - including the genomic scores not including HLA variants (Sig_score, Sigstrong_score, Sigstrongweak_score) and those with the integration of HLA regions (Sig_hla, Sigstrong_hla, Sigstrongweak_hla) - showed statistically

significant differences between the positive and negative progression groups (P > 0.05). The mean scores between the two groups are nearly identical across all clinical phenotypes considered, demonstrating that these scores have poor capability in discriminating the course of disability.

7. Survival analysis

Associations between genetic variables and time to-event outcomes were evaluated through univariate survival analyses and survival analyses covaried by sex and age at the disease onset. Patients have been dichotomized basing on their genotype status (positive/negative for the risk allele) and cox survival curves have been constructed using time to event in days. For the PRSs, patients have been stratified in risk quartiles, basing on the PRS value, and we compared patients belonging to the lower quartile (25% lower risk) with those belonging to the higher quartiles (25% higher risk).

7.1. HLA-A*02:01

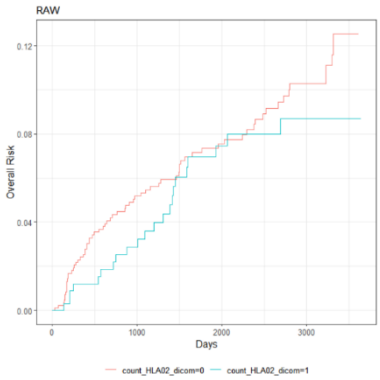

Figure 11. Kaplan-Meier cumulative risk curve generated from RAW data, stratified by HLA-A2 status. The plot illustrates the overall risk over time (days) for HLA-A2 negative patients (red line, count_HLA02_dicom=0) and HLA-A2 positive patients (blue line, count_HLA02_dicom=1).

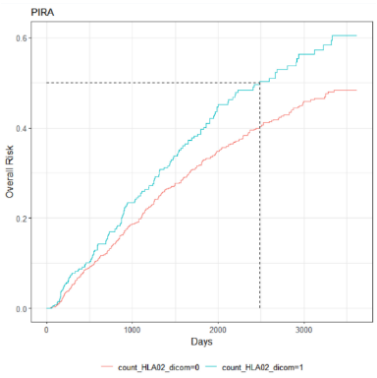

Figure 12. Kaplan-Meier cumulative risk curve generated from PIRA data, stratified by HLA-A2 status. The plot illustrates the overall risk over time (days) for HLA-A2 negative patients (red line, count_HLA02_dicom=0) and HLA-A2 positive patients (blue line, count_HLA02_dicom=1).

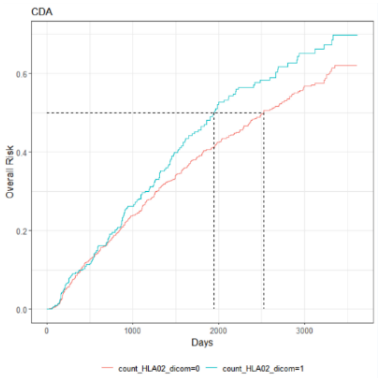

Figure 13. Kaplan-Meier cumulative risk curve generated from CDA data, stratified by HLA-A2 status. The plot illustrates the overall risk over time (days) for HLA-A2 negative patients (red line, count_HLA02_dicom=0) and HLA-A2 positive patients (blue line, count_HLA02_dicom=1).

UNIVARIATE	COVARIATE (SEX & AGE OF ONSET)
P=0.288	P=0.3858
HR=0.7717	HR=0.80824

UNIVARIATE	COVARIATE (SEX & AGE OF ONSET)
P=0.00402	P=0.00538
HR=1.3398	HR=1.329797

UNIVARIATE	COVARIATE (SEX & AGE OF ONSET)
P=0.0309	P=0.0311
HR=1.22394	HR=1.225486

The allele HLA-A*02 was significantly associated with a shorter time to PIRA and CDA event (**Figures 12 and 13**), but not with RAW event (**Figure 11**). For PIRA (**Figure 12**), this increased risk was highly significant in both the univariate model (P = 0.00402, HR = 1.3398) and after adjusting for sex and age of onset (P = 0.00538, HR = 1.329797). Similarly, CDA (**Figure 13**)

confirmed this detrimental trend with statistical significance (P = 0.0309, HR = 1.22394 for univariate; P = 0.0311, HR = 1.225486 for adjusted).

7.2. HLA-DRB1*15:01

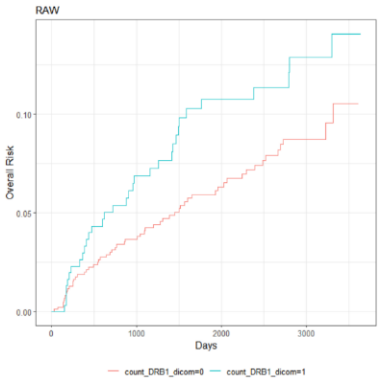

Figure 14. Kaplan-Meier cumulative risk curve generated from RAW data, stratified by HLA-DR15 status. The plot illustrates the overall risk over time (days) for HLA-DR15 negative patients (red line, count_DRB1_dicom=0) and HLA-DR15 positive patients (blue line, count_DRB1_dicom=1).

UNIVARIATE
P=0.0559
HR=1.5135

COVARIATE (SEX & AGE OF ONSET)
P=0.105
HR=1.42967

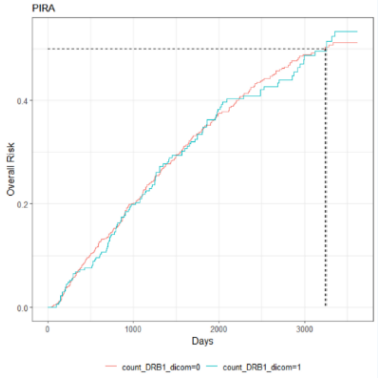

Figure 15. Kaplan-Meier cumulative risk curve generated from PIRA data, stratified by HLA-DR15 status. The plot illustrates the overall risk over time (days) for HLA-DR15 negative patients (red line, count_DRB1_dicom=0) and HLA-DR15 positive patients (blue line, count_DRB1_dicom=1).

UNIVARIATE
P=0.99
HR=1.00134

COVARIATE (SEX & AGE OF ONSET)
P=0.985
HR=1.00201

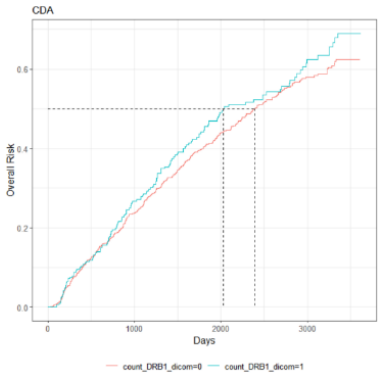

Figure 16. Kaplan-Meier cumulative risk curve generated from CDA data, stratified by HLA-DR15 status. The plot illustrates the overall risk over time (days) for HLA-DR15 negative patients (red line, count_DRB1_dicom=0) and HLA-DR15 positive patients (blue line, count_DRB1_dicom=1).

UNIVARIATE
P=0.224
HR=1.12287

COVARIATE (SEX & AGE OF ONSET)
P=0.2691
HR=1.11203

HLA-DR15 positive patients demonstrated a trend for a shorter time to reach RAW event (**Figure 14**), that was border-line significant in the univariate model (P = 0.0559, HR = 1.5135), though this trend weakened when adjusting for sex and age of onset (P = 0.105, HR = 1.42967). In contrast, no association was observed for PIRA and CDA (**Figures 15 and 16**).

7.3.APOE ε4

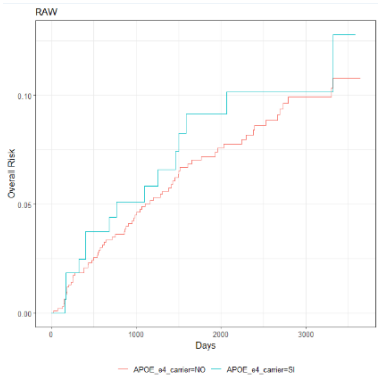

Figure 17. Kaplan-Meier cumulative risk curve generated from RAW data, stratified by APOE status. The plot illustrates the overall risk over time (days) for APOE e4 negative patients (red line, APOE_e4_carrier=NO) and APOE e4 positive patients (blue line, APOE_e4_carrier=SI).

UNIVARIATE
P=0.59
HR=1.2654

COVARIATE (SEX & AGE OF ONSET)
P=0.6429
HR=1.14160

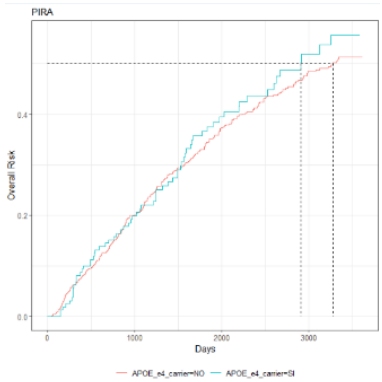

Figure 18. Kaplan-Meier cumulative risk curve generated from PIRA data, stratified by APOE status. The plot illustrates the overall risk over time (days) for APOE e4 negative patients (red line, APOE_e4_carrier=NO) and APOE e4 positive patients (blue line, APOE_e4_carrier=SI).

UNIVARIATE
P=0.573
HR=1.07826

COVARIATE (SEX & AGE OF ONSET)
P=0.532
HR=1.087489

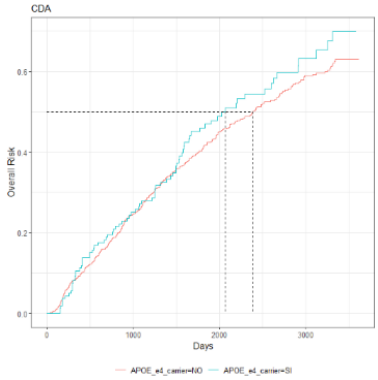

Figure 19. Kaplan-Meier cumulative risk curve generated from CDA data, stratified by APOE status. The plot illustrates the overall risk over time (days) for APOE e4 negative patients (red line, APOE_e4_carrier=NO) and APOE e4 positive patients (blue line, APOE_e4_carrier=SI).

UNIVARIATE
P=0.331
HR=1.1247

COVARIATE (SEX & AGE OF ONSET)
P=0.324
HR=1.12721

The analysis of survival curves stratified by APOE ε4 carrier status revealed a consistent neutral pattern across all three disability outcomes, showing no statistically significant impact on overall risk (**Figures 17, 18 and 19**).

7.4. rs10191329 (DYSF-ZNF638 locus)

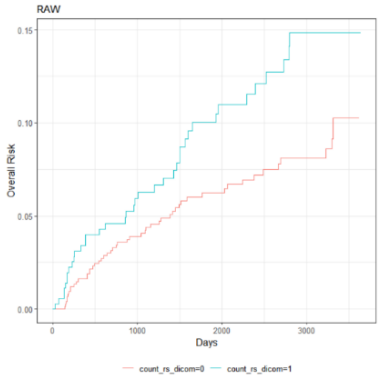

Figure 20. Kaplan-Meier cumulative risk curve generated from RAW data, stratified by rs10191329 status. The plot illustrates the overall risk over time (days) for rs10191329 patients homozygous for the protective allele (red line, count_rs_dicom=0) and patients' carriers of the rs10191329 risk allele (blue line, count_rs_dicom=1).

UNIVARIATE
P=0.0243
HR=1.6174

COVARIATE (SEX & AGE OF ONSET)
P=0.0190
HR=1.66119

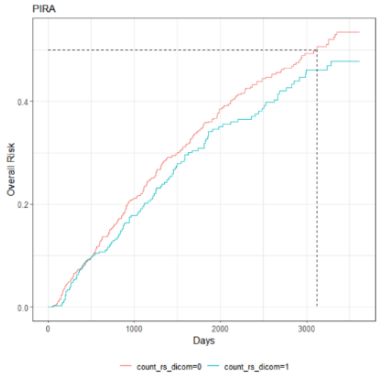

Figure 21. Kaplan-Meier cumulative risk curve generated from PIRA data, stratified by rs10191329 status. The plot illustrates the overall risk over time (days) for rs10191329 patients homozygous for the protective allele (red line, count_rs_dicom=0) and patients' carriers of the rs10191329 risk allele (blue line, count_rs_dicom=1).

UNIVARIATE
P=0.16
HR=0.8598

COVARIATE (SEX & AGE OF ONSET)
P=0.108
HR=0.83936

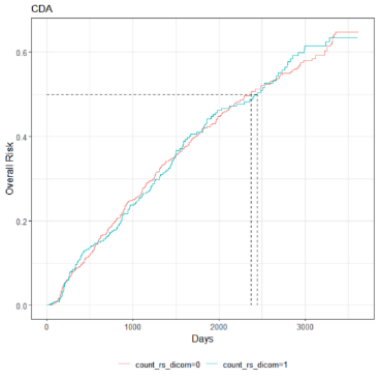

Figure 22. Kaplan-Meier cumulative risk curve generated from CDA data, stratified by rs10191329 status. The plot illustrates the overall risk over time (days) for rs10191329 patients homozygous for the protective allele (red line, count_rs_dicom=0) and patients' carriers of the rs10191329 risk allele (blue line, count_rs_dicom=1).

UNIVARIATE
P=0.968
HR=0.99613

COVARIATE (SEX & AGE OF ONSET)
P=0.865
HR=0.983709

The risk allele (A) of the severity associated SNP rs10191329 was nominally associated with a shorter time to reach RAW (**Figure 20**) event in both the univariate model (P = 0.0243, HR = 1.6174) and when adjusting for sex and age of onset (P = 0.0190, HR = 1.66119). No difference was observed for PIRA and CDA events (**Fires 21 and 22**).

7.5. PRSs

7.5.1. Sig_score

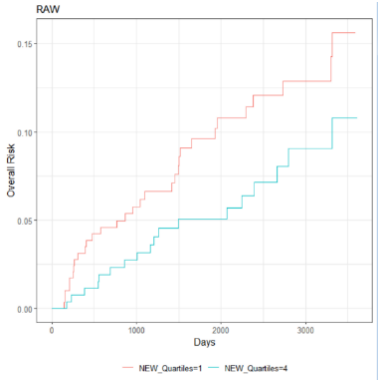

Figure 23. Cumulative risk analysis by PRS sig_score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using RAW data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sig_score.

UNIVARIATE
P=0.0932
HR=0.8472

COVARIATE (SEX & AGE OF ONSET)
P=0.0735
HR=0.83839

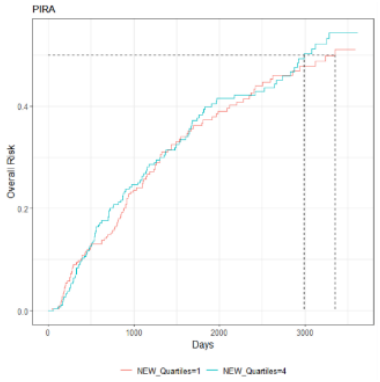

Figure 24. Cumulative risk analysis by PRS sig_score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using PIRA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sig_score.

UNIVARIATE
P=0.683
HR= 1.01843

COVARIATE (SEX & AGE OF ONSET)
P=0.563
HR=1.026383

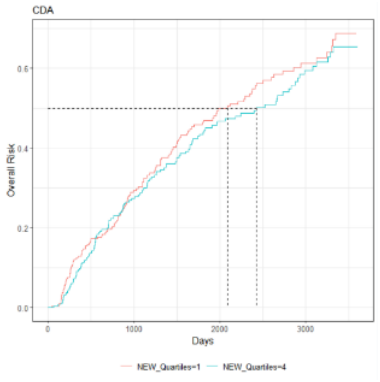

Figure 25. Cumulative risk analysis by PRS sig_score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using CDA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sig_score.

UNIVARIATE
P=0.441
HR=0.96925

COVARIATE (SEX & AGE OF ONSET)
P=0.547
HR=0.975791

7.5.2. Sig_hla

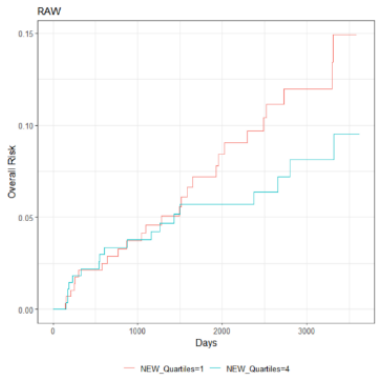

Figure 26. Cumulative risk analysis by PRS sig_hla score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using RAW data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sig_hla.

UNIVARIATE
P=0.219
HR=0.8818

COVARIATE (SEX & AGE OF ONSET)
P=0.9027
HR=1.012283

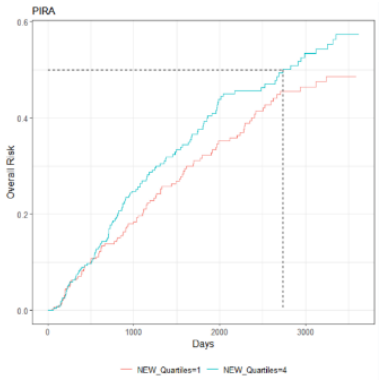

Figure 27. Cumulative risk analysis by PRS sig_hla score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using PIRA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sig_hla.

UNIVARIATE
P=0.102
HR=1.07649

COVARIATE (SEX & AGE OF ONSET)
P=0.0785
HR=1.086512

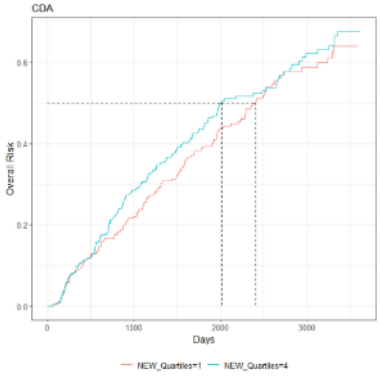

Figure 28. Cumulative risk analysis by PRS sig_hla score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using CDA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sig_hla.

UNIVARIATE
P=0.318
HR=1.04188

COVARIATE (SEX & AGE OF ONSET)
P=0.0585
HR=1.083997

7.5.3. Sigstrong_score

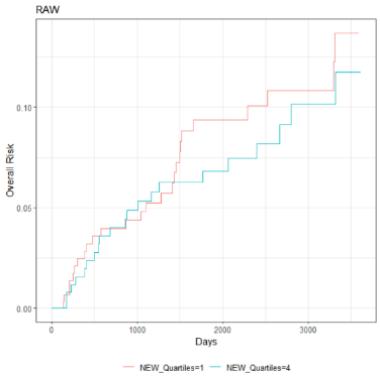

Figure 29. Cumulative risk analysis by PRS sigstrong_score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using RAW data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrong_score.

UNIVARIATE
P=0.586
HR=0.94816

COVARIATE (SEX & AGE OF ONSET)
P= 0.7518
HR=0.969902

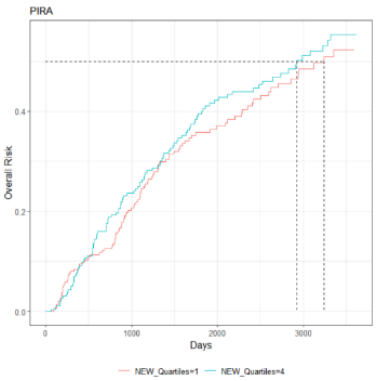

Figure 30. Cumulative risk analysis by PRS sigstrong_score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using PIRA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrong_score.

UNIVARIATE
P=0.481
HR=1.03240

COVARIATE (SEX & AGE OF ONSET)
P= 0.3901
HR=1.040291

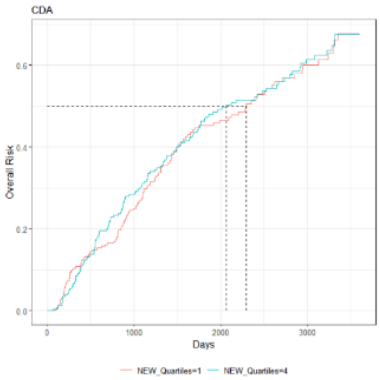

Figure 31. Cumulative risk analysis by PRS sigstrong_score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using CDA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrong_score.

UNIVARIATE
P=0.771
HR=1.01202

COVARIATE (SEX & AGE OF ONSET)
P= 0.578
HR=01.023365

7.5.4. Sigstrong_hla

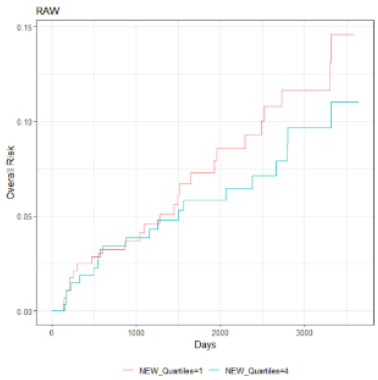

Figure 32. Cumulative risk analysis by PRS sigstrong_hla score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using RAW data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrong_hla score.

UNIVARIATE
P=0.403
HR=0.91970

COVARIATE (SEX & AGE OF ONSET)
P= 0.554
HR=0.94374

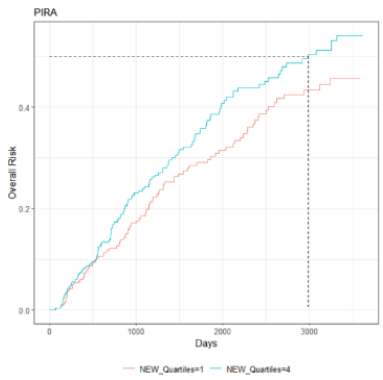

Figure 33. Cumulative risk analysis by PRS sigstrong_hla score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using PIRA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrong_hla score.

UNIVARIATE
P=0.101
HR=1.08002

COVARIATE (SEX & AGE OF ONSET)
P= 0.104
HR=1.079652

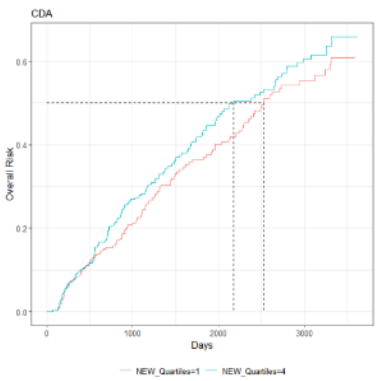

Figure 34. Cumulative risk analysis by PRS sigstrong_hla score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using CDA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrong_hla score.

UNIVARIATE
P=0.239
HR=1.05108

COVARIATE (SEX & AGE OF ONSET)
P= 0.197
HR=1.056220

7.5.5. Sigstrongweak_score

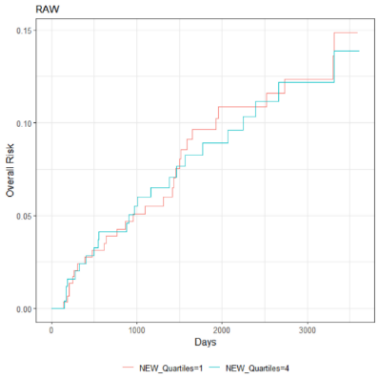

Figure 35. Cumulative risk analysis by PRS sigstrongweak_score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using RAW data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrongweak_score.

UNIVARIATE
P=0.895
HR=0.98792

COVARIATE (SEX & AGE OF ONSET)
P= 0.877
HR=0.98549

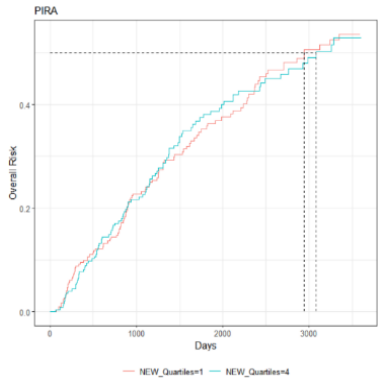

Figure 36. Cumulative risk analysis by PRS sigstrongweak_score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using PIRA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrongweak_score.

UNIVARIATE
P=0.988
HR=0.9993256

COVARIATE (SEX & AGE OF ONSET)
P= 0.838
HR=1.009507

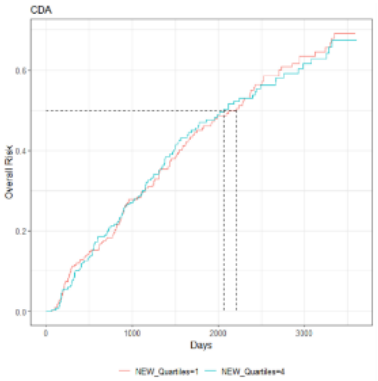

Figure 37. Cumulative risk analysis by PRS sigstrongweak_score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using CDA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrongweak_score.

UNIVARIATE
P=0.926
HR=0.996198

COVARIATE (SEX & AGE OF ONSET)
P= 0.917
HR=1.004323

7.5.6. Sigstrongweak_hla

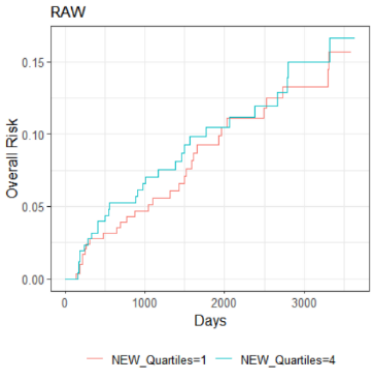

Figure 38. Cumulative risk analysis by PRS sigstrongweak_hla score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using RAW data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrongweak_hla score.

UNIVARIATE
P=0.684
HR=1.03597

COVARIATE (SEX & AGE OF ONSET)
P= 0.630
HR=1.04360

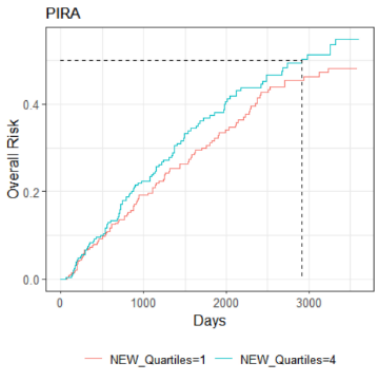

Figure 39. Cumulative risk analysis by PRS sigstrongweak_hla score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using PIRA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrongweak_hla score.

UNIVARIATE
P=0.225
HR=1.05796

COVARIATE (SEX & AGE OF ONSET)
P= 0.386
HR=1.041364

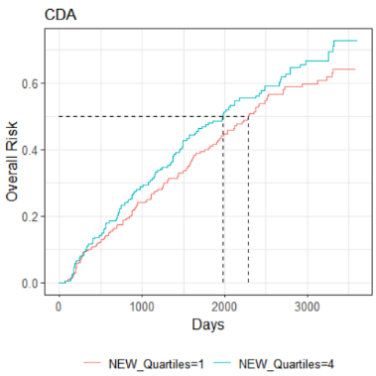

Figure 40. Cumulative risk analysis by PRS sigstrongweak_hla score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using CDA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) sigstrongweak_hla score.

UNIVARIATE
P=0.116
HR=1.06661

COVARIATE (SEX & AGE OF ONSET)
P= 0.185
HR=1.056340

We compared patients with the higher risk given by PRS (quartile 4) with patient given by lower risk (quartile 1). For the scores including HLA markers, the visual analysis of survival curves shows a trend for patients belonging to the high risk category to reach PIRA and CDA before patients belonging to the low risk category. This does not reach the statistical significance, but we observe a borderline significance in the case of the score including only HLA variants (SIG-hla) when covariates are added to the analysis: (PIRA: p=0,785; CDA p=0,585).

Conversely, all the scores based only on non-HLA variants (Sig_score, , Sigstrong_score, Sigstrongweak_score,) – revealed a broadly uniform lack of statistically significant impact on overall risk when comparing the lowest and highest quartiles (Q1 vs. Q4), independently from the stringency applied to the filtered metrics.

7.5.7. Severity score

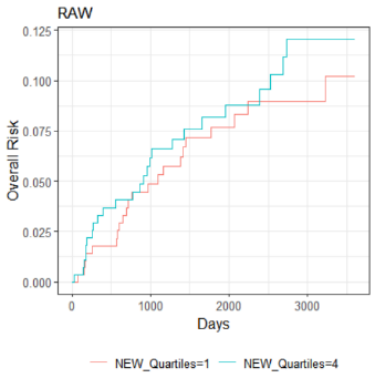

Figure 41. Cumulative risk analysis by PRS severity score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using RAW data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) severity score.

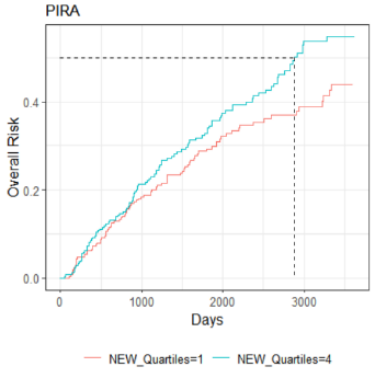

Figure 42. Cumulative risk analysis by PRS severity score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using PIRA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) severity score.

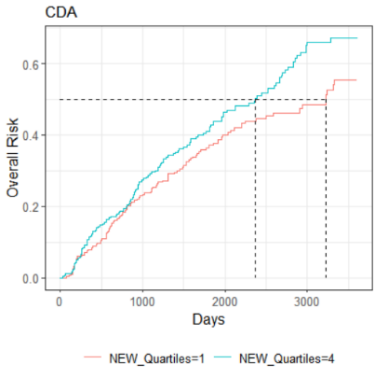

Figure 43. Cumulative risk analysis by PRS severity score quartiles. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using CDA data. The analysis compares patient cohorts stratified into the lowest quartile (Q1, red line) and highest quartile (Q4, blue line) of the polygenic risk score (PRS) severity score.

UNIVARIATE	COVARIATE	UNIVARIATE	COVARIATE	UNIVARIATE	COVARIATE
P=0.537	(SEX & AGE OF ONSET)	P=0.0686	(SEX & AGE OF ONSET)	P=0.0397	(SEX & AGE OF ONSET)
HR=1.06208	P= 0.372	HR=1.9012	P= 0.0852	HR=1.09162	P= 0.0326
	HR=1.09514		HR=1.086088		HR=1.097019

For the severity score, with the univariate model we observed a nominally significant association with a shorter time to reach CDA (P=0.0397) in patients presenting with high risk score compared with those with low risk, that maintained significance also when applying the correction for sex and age at onset (**Figure 43**). Furthermore, a similar border-line significance trend was observed also with for a shorter time to reach PIRA event (**Figure 42**).

8. Smoke

8.1.Association between cigarette smoking and clinical phenotypes

	RAW	PIRA	CDA
OR (95% CI)	1,0368 (0,6237-1,7137)	1,3083 (0,9749-1,7556)	1,2922 (0,9762-1,7106)
P-value	0,8878	0,0734	0,0732

Table 14. Association between smoke and clinical disability progression phenotypes. The table presents the results of a univariate statistical analysis evaluating the impact of cigarette smoking on Relapse-Associated Worsening (RAW), Progression Independent of Relapse Activity (PIRA), and Confirmed Disability Accumulation (CDA).

8.2. Survival analyses

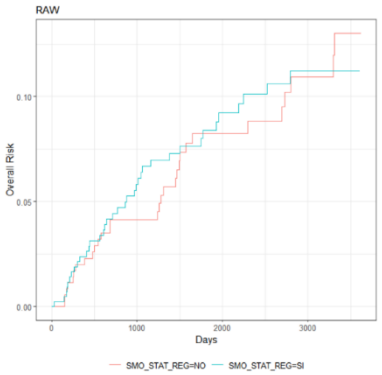

Figure 44. Cumulative risk analysis based on cigarette smoking status. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using the RAW phenotype workflow. Patients are stratified into non-smokers (red line) and cigarette smoking patients (blue line).

UNIVARIATE
P=0.929
HR=1.02218

COVARIATE (SEX & AGE OF ONSET)
P= 0.935
HR=0.97976

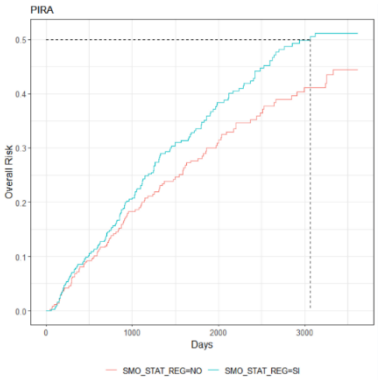

Figure 45. Cumulative risk analysis based on cigarette smoking status. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using the PIRA phenotype workflow. Patients are stratified into non-smokers (red line) and cigarette smoking patients (blue line).

UNIVARIATE
P=0.074
HR=1.2463

COVARIATE (SEX & AGE OF ONSET)
P= 0.1195
HR=1.214871

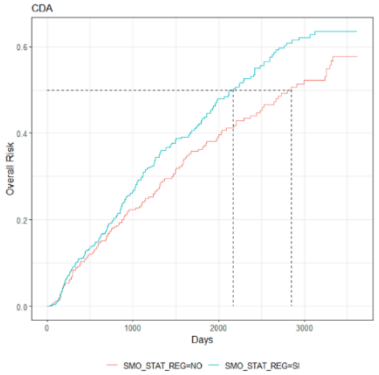

Figure 46. Cumulative risk analysis based on cigarette smoking status. Kaplan-Meier curves displaying the overall risk over a follow-up period of days, generated using the CDA phenotype workflow. Patients are stratified into non-smokers (red line) and cigarette smoking patients (blue line).

UNIVARIATE
P=0.0425
HR=1.2504

COVARIATE (SEX & AGE OF ONSET)
P= 0.0816
HR=1.215171

We observed a border-line significance for the association between smoke and PIRA or CDA event (P = 0.073). This data parallels what observed with the survival curves, in which smokers reach both PIRA and CDA earlier compared to non-smokers (**Figures 45 and 46**). This trend is nominally statistically significant for CDA (P = 0.0425, HR = 1.2504). Together, these results show that while cigarette smoking does not affect acute relapses, it may act as a driver of underlying, long-term progression.

Discussion

This study investigated the ability of individual genetic variables, polygenic risk scores, and the environmental factor of cigarette smoking to predict long-term disability accumulation in a large cohort of approximately 1,500 multiple sclerosis (MS) patients.

Since the sample set is not powered for an unbiased association analysis on all the genome, we decided to select a number of HLA and not-HLA genetic variants already reported for association with susceptibility to MS or with disease severity, and to construct Polygenic Risk Scores (PRS) based on known susceptibility or severity associated genomic variations.

In particular, we selected two well-known HLA variants that are strongly associated with MS susceptibility (HLA-DRB1*15 and HLA-A*02), two non-HLA genetic variants that in literature have been reported to be associated with MS severity (rs10191329, an intergenic SNP located in the DYSF-ZNF638 genomic region , reported to be associated at GWAS level with a clinical and prognostic score ((Age-Related Multiple Sclerosis Severity Score)[81],and APOE ε4 allele that has been recently reported to be associated with relapse-associated worsening in a small sample set of patients with relapsing-remitting MS. Furthermore, we constructed 7 PRS: 6 with different combination of non-HLA SNPs and HLA variants associated with MS susceptibility and one PRS including genomic variants associated with disease severity.

We focused our analysis on three disability outcomes: relapse-associated worsening (RAW), progression independent of relapse activity (PIRA), and confirmed disability accumulation (CDA). Two complementary analytical approaches were applied: a basic association analysis evaluating the risk for developing each endpoint, and Kaplan–Meier survival analysis with Cox proportional hazards models, both in univariate form and after adjustment for sex and age of onset. Comparing the results of these two approaches proved informative in several cases, as associations identified in one framework were not always confirmed by the other.

HLA-A*02:01 allele is a well-known protective allele for MS susceptibility, but its role in disease progression is less clear. In our cohort, HLA-A*02:01 allele was significantly associated with a shorter time to PIRA and CDA event and with a higher risk to reach PIRA event. These finding may look unexpected, as HLA-A*02:01 is generally considered protective against developing MS.

One possible explanation is that the protective effect of HLA-A*02:01 is mainly related to disease onset, where it may help regulate peripheral immune responses, for example by improving the clearance of autoreactive immune cells or EBV-infected B cells. Once MS has developed, however, this protective effect may no longer influence the chronic

neurodegenerative processes that drive PIRA. This is consistent with the current understanding that PIRA is largely driven by mechanisms within the central nervous system, such as chronic inflammation, microglial activation, and axonal loss, rather than by peripheral immune activity[82]. Overall, these findings suggest that the genetic factors influencing MS susceptibility may differ from those involved in disease progression. Further studies are needed to confirm the role of HLA-A*02:01 and other HLA class I alleles in disability accumulation.

HLA-DRB1*15:01 represents the strongest and most consistently reported genetic risk factor for MS susceptibility. However, whether it also influence disease progression and long-term disability is still debated. Our results reflect this uncertainty.

Carrying HLA-DRB1*15:01 allele was associated with a higher risk of RAW (OR 1.64, P = 0.029) and CDA (OR 1.33, P = 0.026), but not PIRA. However, these findings were not confirmed in the survival analysis. For RAW, there was only a borderline association with a shorter time to reach RAW event in the in the unadjusted survival analysis (HR 1.52, P = 0.056), which became non-significant after adjusting for sex and age at disease onset (P = 0.105).

These findings suggest that the associations seen in analysis were likely influenced by stratification in age at onset and sex between carriers and non-carriers, rather than by an independent effect of HLA-DRB1*15:01 itself. Once these factors were considered, the allele was no longer associated with any of the disability outcomes. This is in line with previous studies, which have reported inconsistent results regarding the role of HLA-DRB1*15:01 in MS progression. Although HLA class II molecules play an important role in relapse-related inflammation, their contribution to long-term disability remains unclear[83]. The lack of association with PIRA is particularly consistent with the idea that progression independent of relapse activity is mainly driven by mechanisms within the central nervous system rather than by peripheral adaptive immune responses[82]. Overall, our findings suggest that, while HLA-DRB1*15:01 is highly relevant for MS susceptibility, it does not appear to be an independent predictor of disability accumulation after accounting for important clinical factors such as age at onset and sex.

The APOE ε4 allele is the most important common genetic risk factor for late-onset Alzheimer's disease and has been shown to influence neuroinflammatory pathways relevant to MS[84]. A recent longitudinal study reported a significant association between APOE ε4 and in a small (n=160) sample set of Italian patients with relapsing MS, suggesting a role for this allele in modulating inflammatory disability accrual[54]. In the present analysis, however, no statistically

significant associations were detected between APOE ε4 carrier status and any of the three disability endpoints. This data, in contrast with the recent literature finding may be due to the fact that our sample set is much bigger and thus we may provide a more precise statistical observation.

These null findings do not exclude a genuine biological role for APOE ε4 in MS progression, but they do suggest caution in generalising results from smaller or selected cohorts. Replication in larger, well-characterised longitudinal cohorts with detailed treatment histories will be necessary to clarify the prognostic value of this variant in MS.

The variant rs10191329, located within the DYSF-ZNF638 locus, was the first genome-wide significant genetic variant associated with MS severity rather than disease susceptibility[53]. In our cohort, this variant was significantly associated with RAW (OR 1.70, P = 0.017), suggesting that carriers have a higher risk of disability accumulation related to relapse-associated inflammatory activity. In contrast, no significant associations were found with PIRA or CDA.

These results partially agree with the original description of rs10191329 as a severity-associated variant, but they also provide a different perspective. The original study linked this variant to faster disability progression and shorter time to requiring a walking aid, proposing that its effect was mainly related to cortical neurodegeneration rather than peripheral immune mechanisms[81]. Therefore, the association with RAW observed in our study, rather than with PIRA, was somewhat unexpected. Our findings should also be considered alongside later replication studies. Campagna et al.[85] were unable to replicate the association between rs10191329 and longitudinal MS severity in an independent MSBase cohort, suggesting that the effect of this variant may be smaller or more dependent on the characteristic of the study population and the definition of the outcome. Our results do not necessarily conflict with these findings. Instead, they suggest that the effect of rs10191329 may be easier to detect when disability accumulation is divided into its different components, rather than analysed as single measure. The association with RAW, but not PIRA, raises the possibility that this variant mainly influences recovery after inflammatory relapses, an effect that could be missed when using broader measures of disability progression.

Overall, these findings suggest that the impact of rs10191329 on MS progression may depend on the phenotype being studied. Larger longitudinal studies using biologically meaningful outcomes such us RAW and PIRA will be important to better understand the role of this variant in disability accumulation.

The analyses performed in this study showed that susceptibility-based polygenic risk scores (PRS) did not predict disability accumulation. None of the six susceptibility-based PRSs tested (Sig_score, Sig_hla, Sigstrong_score, Sigstrong_hla, Sigstrongweak_score, and Sigstrongweak_hla) showed significant associations with RAW, PIRA, or CDA, either when comparing individuals in the highest and lowest PRS quartiles or when comparing patients who did and did not experience each disability outcome.

The score including only the 32 HLA variants from the 2019 IMSGC GWAS[52] (Sig_hla) shows a border-line association that however does not reach statistical significance with time to-event in case of PIRA and CDA, with patients belonging to the high risk group reaching the disability event earlier than those belonging to the low risk, but this trend disappeared when non-HLA SNPs from the 2019 IMSGC GWAS[52] were incorporated. The result on the Sig_hla is probably partially driven by the effect of HLA-DRB1*15 and HLA-A*02 alleles, that are included in the score.

These findings support the growing evidence that the genetic factors influencing MS susceptibility are different from those driving disease progression. Most variants identified by susceptibility GWAS are involved in immune processes such as antigen presentation, lymphocyte activation, and cytokine signalling, which determine the risk of developing MS. In contrast, disability progression, particularly PIRA, is thought to be driven mainly by mechanisms within the central nervous system, including microglial activation, synaptic loss, mitochondrial dysfunction, and iron accumulation. These processes are unlikely to be captured by susceptibility-based PRS.

These results may have potential implications for the interpretation of PRSs in MS. Although susceptibility-based PRSs may be useful for estimating the risk of developing MS, they showed limited ability to predict disease progression or disability accumulation in the present cohort. These observations suggest that future studies could further investigate progression-specific PRSs based on genetic variants associated with disability outcomes rather than disease susceptibility.

In contrast to the susceptibility-based scores, the Severity Score showed a significant association with CDA ($P = 0.040$) and a borderline association with PIRA ($P = 0.069$), while no association was found with RAW ($P = 0.54$). This pattern suggests that the Severity Score captures genetic factors related to disability progression rather than relapse-associated worsening. Unlike susceptibility scores, it is based on genetic variants associated with disease

severity, which are thought to reflect mechanisms involved in neurodegeneration and central nervous system resilience[81].

The association with PIRA, but not RAW, is consistent with the current understanding that progression independent of relapse activity is mainly driven by processes within the central nervous system, such as chronic microglial activation, oxidative stress, mitochondrial dysfunction, and reduced neuroprotective capacity[82]. These mechanisms are more likely to be influenced by severity-associated genetic variants than by the immune-related variants identified in susceptibility GWAS.

Overall, these findings suggest that severity-based polygenic scores may be more useful than susceptibility-based scores for identifying patients at greater risk of long-term disability accumulation, particularly disability that develops independently of relapses. However, these results should be validated in larger, independent cohorts.

Cigarette smoking is one of the best-established modifiable risk factors for MS and has been linked to both an increased risk of developing the disease and worse clinical outcomes. In our cohort, smoking was significantly associated with CDA (OR 1.29, P = 0.043) and showed a borderline association with PIRA (OR 1.31, P = 0.073), while no association was found with RAW. This pattern suggests that smoking may mainly contribute to disability through mechanisms involved in gradual disease progression rather than by increasing relapse-related disability.

These findings are consistent with previous studies showing that smoking may accelerate neurodegeneration through several mechanisms, including oxidative stress, chronic inflammation, increased production of pro-inflammatory cytokines, blood-brain barrier dysfunction, and tissue hypoxia[86]. These processes are thought to contribute to the chronic neurodegenerative changes that underlie PIRA rather than to the acute inflammatory activity responsible for relapses.

From a clinical perspective, these results further support the importance of smoking cessation as part of MS management. Smoking appears to influence not only the risk of developing MS but also long-term disability after disease onset.

Conclusions

This study suggests that the genetic factors influencing MS susceptibility are different from those involved in disease progression. Susceptibility-based polygenic risk scores were not

associated with disability outcomes, whereas the Severity Score showed associations with PIRA and CDA, supporting its potential as a marker of disease progression. Among the individual genetic variants, rs10191329 was associated with RAW, while HLA-A*02:01 was associated with PIRA and CDA. In contrast, HLA-DRB1*15:01 and APOE ε4 were not associated with disability after adjustment for confounding factors. Smoking was associated with CDA, highlighting the importance of modifiable environmental factors in long-term disability. Overall, these findings support the need for progression-specific genetic markers to improve prognosis and move towards more personalised management of MS.

BIBLIOGRAPHY

1. Correale, J., Gaitán, M. I., Ysrraelit, M. C. & Fiol, M. P. Progressive multiple sclerosis: from pathogenic mechanisms to treatment. *Brain J. Neurol.* **140**, 527–546 (2017).

2. Portaccio, E. *et al.* Multiple sclerosis: emerging epidemiological trends and redefining the clinical course. *Lancet Reg. Health - Eur.* **44**, 100977 (2024).

3. Sclerosi Multipla: età media della diagnosi, forma prevalente e numeri nel mondo - Disabili.com. https://www.disabili.com/medicina/articoli-qmedicinaq/sclerosi-multipla-eta-media-della-diagnosi-forma-prevalente-e-numeri-nel-mondo?highlight=WyJhaSIsNjksImFpIDY5Il0=.

4. Montague, T., Drummond, J., Ng, K. & Parratt, J. Advancements in multiple sclerosis. *Intern. Med. J.* **55**, 895–904 (2025).

5. Hafler, D. A. Multiple sclerosis. *J. Clin. Invest.* **113**, 788–794 (2004).

6. Dighriri, I. M. *et al.* An Overview of the History, Pathophysiology, and Pharmacological Interventions of Multiple Sclerosis. *Cureus* **15**, e33242.

7. Portaccio, E. *et al.* Multiple sclerosis: emerging epidemiological trends and redefining the clinical course. *Lancet Reg. Health - Eur.* **44**, 100977 (2024).

8. Lublin, F. D. & Reingold, S. C. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. *Neurology* **46**, 907–911 (1996).

9. Lublin, F. D. *et al.* Defining the clinical course of multiple sclerosis. *Neurology* **83**, 278–286 (2014).

10. Klineova, S. & Lublin, F. D. Clinical Course of Multiple Sclerosis. *Cold Spring Harb. Perspect. Med.* **8**, a028928 (2018).

11. Lauerer, M. *et al.* Predictors of early disability accumulation in newly diagnosed multiple sclerosis: clinical, imaging and cerebrospinal fluid measures. *J. Neurol. Neurosurg. Psychiatry* **96**, 900–907 (2025).

12. Müller, J. *et al.* Harmonizing Definitions for Progression Independent of Relapse Activity in Multiple Sclerosis: A Systematic Review. *JAMA Neurol.* **80**, 1232–1245 (2023).

13. Sharma, R. McDonald diagnostic criteria for multiple sclerosis | Radiology Reference Article | Radiopaedia.org. *Radiopaedia* https://radiopaedia.org/articles/mcdonald-diagnostic-criteria-for-multiple-sclerosis-4 doi:10.53347/rID-57578.

14. Montalban, X. *et al.* Diagnosis of multiple sclerosis: 2024 revisions of the McDonald criteria. *Lancet Neurol.* **24**, 850–865 (2025).

15. Mey, G. M., Mahajan, K. R. & DeSilva, T. M. Neurodegeneration in multiple sclerosis. *Wires Mech. Dis.* **15**, e1583 (2023).

16. Boutitah-Benyaich, I. *et al.* Multiple sclerosis: molecular pathogenesis and therapeutic intervention. *Signal Transduct. Target. Ther.* **10**, 324 (2025).

17. Dargahi, N. *et al.* Multiple Sclerosis: Immunopathology and Treatment Update. *Brain Sci.* **7**, 78 (2017).

18. Haase, S. & Linker, R. A. Inflammation in multiple sclerosis. *Ther. Adv. Neurol. Disord.* **14**, 17562864211007687 (2021).

19. Hedegaard, C. J. *et al.* Autoantibodies to myelin basic protein (MBP) in healthy individuals and in patients with multiple sclerosis: a role in regulating cytokine responses to MBP. *Immunology* **128**, e451–e461 (2009).

20. Liu, R. *et al.* Autoreactive lymphocytes in multiple sclerosis: Pathogenesis and treatment target. *Front. Immunol.* **13**, 996469 (2022).

21. Miyazaki, Y. & Niino, M. B-cell depletion therapy for multiple sclerosis. *Immunol. Med.* **45**, 54–62 (2022).

22. Kinzel, S. & Weber, M. S. B Cell-Directed Therapeutics in Multiple Sclerosis: Rationale and Clinical Evidence. *CNS Drugs* **30**, 1137–1148 (2016).

23. Belbasis, L., Bellou, V., Evangelou, E., Ioannidis, J. P. A. & Tzoulaki, I. Environmental risk factors and multiple sclerosis: an umbrella review of systematic reviews and meta-analyses. *Lancet Neurol.* **14**, 263–273 (2015).

24. Olsson, T., Barcellos, L. F. & Alfredsson, L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. *Nat. Rev. Neurol.* **13**, 25–36 (2017).

25. Imani, D., Razi, B., Motallebnezhad, M. & Rezaei, R. Association between vitamin D receptor (VDR) polymorphisms and the risk of multiple sclerosis (MS): an updated meta-analysis. *BMC Neurol.* **19**, 339 (2019).

26. Miclea, A., Bagnoud, M., Chan, A. & Hoepner, R. A Brief Review of the Effects of Vitamin D on Multiple Sclerosis. *Front. Immunol.* **11**, (2020).

27. Carlberg, C. & Campbell, M. J. Vitamin D receptor signaling mechanisms: Integrated actions of a well-defined transcription factor. *Steroids* **78**, 127–136 (2013).

28. Abbasi, H. *et al.* Vitamin D in Multiple Sclerosis: A Comprehensive Umbrella Review. *J. Nutr.* **154**, 3505–3520 (2024).

29. Vitturi, B. K. *et al.* Environmental risk factors for multiple sclerosis: a comprehensive systematic review and meta-analysis. *J. Neurol.* **272**, 513 (2025).

30. Nishanth, K. *et al.* Role of Smoking in the Pathogenesis of Multiple Sclerosis: A Review Article. *Cureus* **12**, (2020).

31. Wang, Z., Xie, J., Wu, C. & Xiao, G. Correlation Between Smoking and Passive Smoking with Multiple Sclerosis and the Underlying Molecular Mechanisms. *Med. Sci. Monit.* **25**, 893–902 (2019).

32. Soldan, S. S. & Lieberman, P. M. Epstein–Barr virus and multiple sclerosis. *Nat. Rev. Microbiol.* **21**, 51–64 (2023).

33. Thorley-Lawson, D. A. EBV Persistence—Introducing the Virus. *Curr. Top. Microbiol. Immunol.* **390**, 151–209 (2015).

34. Handel, A. E. *et al.* An Updated Meta-Analysis of Risk of Multiple Sclerosis following Infectious Mononucleosis. *PLoS ONE* **5**, e12496 (2010).

35. Chen, J., Zhang, S., Zhao, Y., Zheng, F. & Cheng, F. Epstein-Barr virus: biology, pathogenesis and therapy of lymphomas. *Discov. Oncol.* **16**, 2154 (2025).

36. Damania, B., Kenney, S. C. & Raab-Traub, N. Epstein-Barr virus: Biology and clinical disease. *Cell* **185**, 3652–3670 (2022).

37. Bjornevik, K., Münz, C., Cohen, J. I. & Ascherio, A. Epstein–Barr virus as a leading cause of multiple sclerosis: mechanisms and implications. *Nat. Rev. Neurol.* **19**, 160–171 (2023).

38. Modelling the structure of full-length Epstein–Barr virus nuclear antigen 1 | Virus Genes | Springer Nature Link. https://link.springer.com/article/10.1007/s11262-014-1101-9.

39. Hedström, A. K. *et al.* High Levels of Epstein–Barr Virus Nuclear Antigen-1-Specific Antibodies and Infectious Mononucleosis Act Both Independently and Synergistically to Increase Multiple Sclerosis Risk. *Front. Neurol.* **10**, 1368 (2020).

40. Stampanoni Bassi, M. *et al.* Obesity worsens central inflammation and disability in multiple sclerosis. *Mult. Scler. J.* **26**, 1237–1246 (2020).

41. Correale, J. & Marrodan, M. Multiple sclerosis and obesity: The role of adipokines. *Front. Immunol.* **13**, 1038393 (2022).

42. Zhou, X. *et al.* Gut microbiome of multiple sclerosis patients and paired household healthy controls reveal associations with disease risk and course. *Cell* **185**, 3467-3486.e16 (2022).

43. Jangi, S. *et al.* Alterations of the human gut microbiome in multiple sclerosis. *Nat. Commun.* **7**, 12015 (2016).

44. Nylander, A. & Hafler, D. A. Multiple sclerosis. *J. Clin. Invest.* **122**, 1180–1188 (2012).

45. Sahi, N., Ciccarelli, O., Houlden, H. & Chard, D. T. Unlocking Multiple Sclerosis Genetics. *Neurology* **105**, e214141 (2025).

46. Patsopoulos, N. A. Genetics of Multiple Sclerosis: An Overview and New Directions. *Cold Spring Harb. Perspect. Med.* **8**, a028951 (2018).

47. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).

48. Gourraud, P.-A., Harbo, H. F., Hauser, S. L. & Baranzini, S. E. The genetics of multiple sclerosis: An up-to-date review. *Immunol. Rev.* **248**, 87–103 (2012).

49. Pishesha, N., Harmand, T. J. & Ploegh, H. L. A guide to antigen processing and presentation. *Nat. Rev. Immunol.* **22**, 751–764 (2022).

50. Hollenbach, J. A. & Oksenberg, J. R. The Immunogenetics of Multiple Sclerosis: A Comprehensive Review. *J. Autoimmun.* **64**, 13–25 (2015).

51. Brynedal, B. *et al.* HLA-A Confers an HLA-DRB1 Independent Influence on the Risk of Multiple Sclerosis. *PLoS ONE* **2**, e664 (2007).

52. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).

53. Novel genetic locus and CNS resilience modulate multiple sclerosis severity. *Nature* **619**, 323–331 (2023).

54. Portaccio, E. *et al.* Relapse-associated worsening in apolipoprotein E ε4 carriers with relapsing multiple sclerosis. *J. Neurol. Neurosurg. Psychiatry* https://doi.org/10.1136/jnnp-2025-338392 (2026) doi:10.1136/jnnp-2025-338392.

55. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).

56. Genome-Wide Association Studies (GWAS). https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies-GWAS.

57. Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **27**, e1608 (2018).

58. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).

59. Zondervan, K. T. & Cardon, L. R. Designing candidate gene and genome-wide case–control association studies. *Nat. Protoc.* **2**, 2492–2501 (2007).

60. Wang, M. & Xu, S. Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity* **123**, 287–306 (2019).

61. Purcell, S., Cherny, S. S. & Sham, P. C. Genetic Power Calculator: design of linkage andassociation genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–150 (2003).

62. Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *The Lancet* **361**, 598–604 (2003).

63. Ho, C. C., Mun, K.-S. & Naidu, R. SNP array technology: an array of hope in breast cancer research. (2013).

64. You, Q., Yang, X., Peng, Z., Xu, L. & Wang, J. Development and Applications of a High Throughput Genotyping Tool for Polyploid Crops: Single Nucleotide Polymorphism (SNP) Array. *Front. Plant Sci.* **9**, (2018).

65. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

66. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).

67. Hardy-Weinberg Law Of Genetic Equilibrium | Science | Research Starters | EBSCO Research. *EBSCO* https://www.ebsco.com.

68. Minor Allele Frequency (MAF) - Gene Inspector Pro. https://help.gene-inspector.pro/terminology/minor-allele-frequency/.

69. Begum, F., Ghosh, D., Tseng, G. C. & Feingold, E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.* **40**, 3777–3784 (2012).

70. Chang, Y.-H. *et al.* Formal Statistical Replication Analysis in Lung Cancer Genome-Wide Association Studies. *medRxiv* 2025.10.02.25337130 (2025) doi:10.1101/2025.10.02.25337130.

71. Cao, X., Xing, L., He, H. & Zhang, X. Views on GWAS statistical analysis. *Bioinformation* **16**, 393–397 (2020).

72. Adam, Y., Samtal, C., Brandenburg, J., Falola, O. & Adebiyi, E. Performing post-genome-wide association study analysis: overview, challenges and recommendations. *F1000Research* **10**, 1002 (2021).

73. Zeng, P. *et al.* Statistical analysis for genome-wide association study. *J. Biomed. Res.* **29**, 285–297 (2015).

74. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genet.* **9**, e1003348 (2013).

75. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. A guide to performing Polygenic Risk Score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

76. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).

77. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

78. Harris, L. *et al.* Genome-wide association testing beyond SNPs. *Nat. Rev. Genet.* **26**, 156–170 (2025).

79. Bashinskaya, V. V., Kulakova, O. G., Boyko, A. N., Favorov, A. V. & Favorova, O. O. A review of genome-wide association studies for multiple sclerosis: classical and hypothesis-driven approaches. *Hum. Genet.* **134**, 1143–1162 (2015).

80. The International Multiple Sclerosis Genetics Consortium. Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study. *N. Engl. J. Med.* **357**, 851–862 (2007).

81. Harroud, A. *et al.* Locus for severity implicates CNS resilience in progression of multiple sclerosis. *Nature* **619**, 323–331 (2023).

82. Guerra, T. & Iaffaldano, P. A Window into New Insights on Progression Independent of Relapse Activity in Multiple Sclerosis: Role of Therapies and Current Perspective. *Int. J. Mol. Sci.* **26**, 884 (2025).

83. Sahi, N., Ciccarelli, O., Houlden, H. & Chard, D. T. Unlocking Multiple Sclerosis Genetics. *Neurology* **105**, e214141 (2025).

84. De Stefano, N. *et al.* Influence of Apolipoprotein E ϵ4 Genotype on Brain Tissue Integrity in Relapsing-Remitting Multiple Sclerosis. *Arch. Neurol.* **61**, 536–540 (2004).

85. Campagna, M. P. *et al.* No evidence for association between rs10191329 severity locus and longitudinal disease severity in 1813 relapse-onset multiple sclerosis patients from the MSBase registry. *Mult. Scler. J.* **30**, 1216–1220 (2024).

86. Heydarpour, P. *et al.* Smoking and worsening disability in multiple sclerosis: A meta-analysis. *Acta Neurol. Scand.* **138**, 62–69 (2018).

Acknowledgements

I would like to thank Professor Sandra D'Alfonso, Professor Maurizio Leone, Professor Maria Trojano and Dr. Giuseppe Lucisano for their contribution to this initiative.

My deepest acknowledgment goes to my academic advisor, Professor Nadia Barizzone, who guided me through the writing of this work. Professor Lucia Corrado, thank you for your constant supervision and valuable insights.

A special thanks goes to Endri Visha and Alaa Habeeb, my tutors during my laboratory experience. Thank you for everything you have taught me over the past two years, for the constant support and for the laughs.

I would like to thank all the PhD students and seniors from the genetic laboratory: Martina, Fjorilda, Beatrice, Giovanni, and Benedetta. Thank you for making this experience worth remembering.

A huge thank you to my colleagues, Margherita, Letizia, and Asia. I would have been lost without you.

Lastly, I would like to thank my family and friends, for the constant support outside the academic life. You have been my anchor throughout these past two years, and it is thanks to all of you that I am here today.