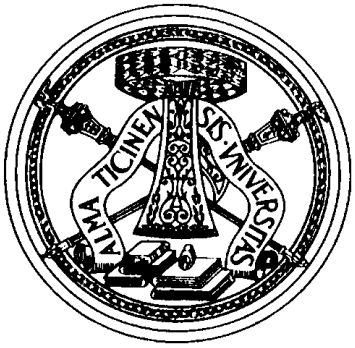
UNIVERSITÀ DEGLI STUDI DI PAVIA

Dottorato di Ricerca in
"*Psicologia, Neuroscienze e Statistica Medica*"



Identification of plasma proteins causally related to Multiple Sclerosis via a Mendelian Randomization approach: a study on multiplex families from the founder population of the Nuoro province (Sardinia)

Ph.D candidate:
Gabriele Morani

Tutor:
Prof.ssa Luisa Bernardinelli

A. A. 2015-2018 – XXXI ciclo

Abstract

Background: The pathogenesis of Multiple Sclerosis (MS) is poorly understood. A better understanding of the causal pathways involved in this disease is needed as a basis for developing new therapies.

Objectives: With this study we try to assess the existence of causal relationships between a large set of candidate plasma proteins and MS. Our analysis is based on 20 multiplex families from the founder and genetically homogeneous population of the Nuoro province, Sardinia (Italy). Our aim is to improve our understanding of the pathophysiological bases of this disease, providing important candidates to be prioritized for further studies on MS and for drug discovery possibly leading to the improvement of the clinical conditions of the subjects affected by this disabling disease.

Methods: We investigated each protein, in turn, for a possible causal effect on MS, taking advantage of the use of Mendelian Randomization (MR) methods to avoid the classical biases that affects observational studies. To overcome the limitations of observational studies we adopted a MR approach to the analysis, where genetic variants act as instrumental variables for the assessment of the putative causal effect. We applied different MR methods based on summary statistics: Inverse-Variance Weighted as the main method and the Weighted Median Estimator and Egger regression for sensitivity analysis purpose. The data supported causality of a number of proteins, which we then checked via bidirectional MR analysis to assess potential reverse causation.

Results: In the end, 3 proteins showed significant results with both Bonferroni and Benjamini-Hochberg corrections, in particular MOBP, ZMYND19 and EFCAB14. Following the bidirectional analysis though, ZMYND19 showed a significant result in the reverse-direction too, suggesting some reverse causation effect. It seems that, in this case, the disease itself could influence the level of this protein in plasma. The final and most interesting findings in the end are therefore MOBP and EFCAB14.

Conclusion: Whereas MR methods are typically applied to high-level exposures, such as obesity and cholesterol, ours is one of the few studies that uses standard MR methods to identify genes that drive the disease by influencing the concentration of their coded proteins, applying a systematic routine of analysis on a very large set of candidate proteins in what seems to be a very promising and

useful exploratory approach. We confirmed two proteins being causally related to MS. The variants in the genes coding for these proteins were found statistically associated to MS in previous studies.

Contents

List of Figures

Figure 15 – KIAA0494 Gene in genomic location: bands according to Ensembl, locations according to GeneLoc.

List of tables

Table 10 – MR methods, betas, standard errors, 95% confidence intervals and uncorrected p-values for zmynd19_hpa020642.

Table 11 – MR methods, betas, standard errors, 95% confidence intervals and uncorrected p-values for kiaa0494_hpa011224.

Table 12 – MR methods, betas, standard errors, 95% confidence intervals and p-values for mobp_hpa035152 (reverse-causation analysis).

Table 13 – MR methods, betas, standard errors, 95% confidence intervals and p-values for zmynd19_hpa020642 (reverse-causation analysis).

Table 14 – MR methods, betas, standard errors, 95% confidence intervals and p-values for kiaa0494_hpa011224 (reverse-causation analysis).

Table 15 – Number of selected Instrumental Variables (IVs) for each protein (antibody ID shown) analysed, ordered higher to lower.

1. Introduction

Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear less.
- Marie Curie

1.1 Presenting the study

Observational epidemiological studies can be subject to a variety of biases and have therefore to face a fundamental problem: it can be very difficult, or even impossible, to separate causal associations from those that arise from confounding or reverse causation.

If a researcher is interested in some disease biomarkers as potential predictors of disease risk, it is not essential that the biomarker-disease association is causal: a demonstrable and consistent association of the biomarker with the disease can be more than sufficient in most of cases. However, if the main interest relies instead in the potential aetiological role of a biomarker possibly modifiable by public health measures or drug treatment, in assessing the impact of a medical intervention, in prioritizing health resources, evidence on a causal association is essential.

Mendelian randomization (MR)[1] is an ensemble of techniques, which have undergone a massive and rapid development in the last few years, in which genetic variants are used to help discern causal from non-causal associations between environmental exposures or biomarkers and disease outcomes. This is made possible by two intrinsic characteristics of genotype: random allocation of parental alleles to zygotes at meiosis, which, being independent of environmental exposures, reduces the potential for confounding in genetic association studies in

a way that resembles randomized treatment allocation in clinical trials[2], and the invariant nature of the DNA sequence and unidirectional flow of biological information, from gene sequence through intermediate phenotypes to disease, which avoids reverse causation[3].

In last years, research has seen a huge increase in MR studies. A systematic review[4] conducted on studies published between 2004 and 2015 revealed that the majority have been in the fields of cardiovascular disease and diabetes (51% of published studies), other disease areas including cancer (10%) and mental health (10%) while most MR studies (86%) have been of disease biomarkers such as blood lipids, body mass index (BMI) or blood pressure, and 50% have used a candidate gene approach to identify suitable instruments. However, the ever-increasing number of genome-wide association studies (GWAS) is now providing a rich source of potential instruments for MR analysis, even though there's still much discussion about proper procedures to use in selecting instruments for this kind of designs[5].

We apply the approach of MR to study Multiple sclerosis (MS).

MS is the most prevalent chronic inflammatory disease of the central nervous system (CNS)[6]. It affects more than 2 million people worldwide, and it is currently incurable. It causes fully or partially reversible episodes of neurologic disability, usually lasting days or weeks. After typically 10 to 20 years, a progressive clinical course develops in many of the persons affected, eventually leading to impaired mobility and cognition.

Currently in the market can be found more than a dozen disease-modifying medications aiming at reducing the frequency of transient episodes of neurologic disability and the accumulation of focal white-matter lesions. Unfortunately, though, no medication fully prevents or reverses the progressive neurologic deterioration caused by the disease, but the question of whether disease-modifying medications can delay clinical progression is controversial.

With this study we try to assess the existence of causal relationships between a large set of candidate plasma proteins and MS, investigating each protein, in turn,

for a possible causal effect on MS, taking advantage of the use of Mendelian Randomization methods to avoid the classical biases that affects observational studies. Our analysis is based on 20 multiplex families from the founder and genetically homogeneous population of the Nuoro province, Sardinia (Italy), which constitutes an interesting choice for the mapping of complex traits, since the structure of isolated populations tends to attenuate the confounding effects of unknown population structure, to show low genetic and environmental heterogeneity and to offer as well simpler underlying association structure.

Our aim is to improve our understanding of the pathophysiological bases of this disease, providing important candidates to be prioritized for further studies on MS and for drug discovery possibly leading to the improvement of the clinical conditions of the subjects affected by this disabling disease.

2. Multiple Sclerosis

In examining disease, we gain wisdom about anatomy and physiology and biology. In examining the person with disease, we gain wisdom about life.

- Oliver Sacks

Multiple Sclerosis (from here abbreviated as MS), also known as "disseminated sclerosis", is primarily an inflammatory demyelinating disease of the CNS first described in 1868 by the French neurologist Jean Martin Charcot. In this disease the fatty myelin sheaths around the axons of the brain and spinal cord are damaged, leading to demyelinated plaque which consist of a well-demarcated hypocellular area characterized by the loss of myelin, relative preservation of axons, and the formation of astrocytic scars[7, 8]. Demyelination is the results of several mechanisms, including immune mediate effects by inflammation cytokines, macrophages or T-cells, as well as an antibody-mediated damage to the myelin and complement-mediated injury (Figure 1), and leads to reduction of conduction speed (saltatory conduction) in the affected nerve, giving rise to clinical symptoms and signals typical of the disease[9].

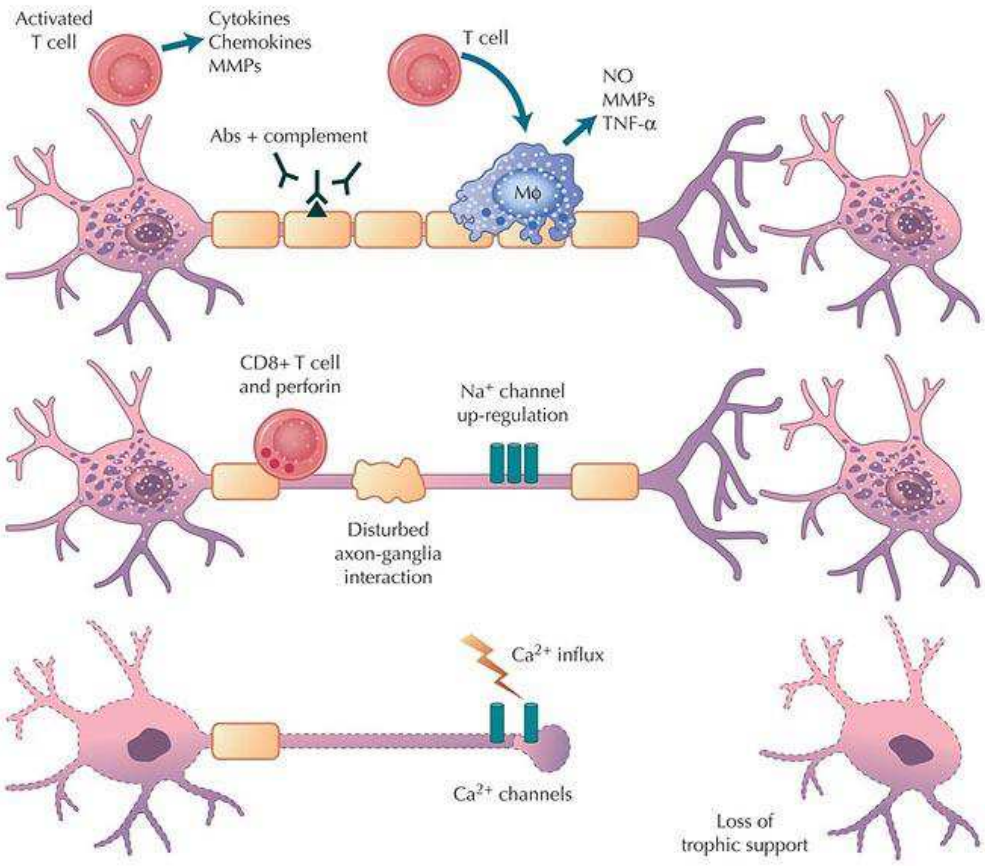


FIGURE 1: **Cascade of events possibly underlying demyelination and axonal degeneration in multiple sclerosis.** Within the central nervous system, activated T lymphocytes release inflammatory cytokines, chemokines, and matrix metalloproteinases (MMPs). Moreover, T cells activate microglia cells/macrophages to enhance phagocytic activity, the production of cytokines, and the release of toxic mediators such as nitric oxide (NO), propagating demyelination and axonal loss. Autoantibodies (Abs) crossing the blood-brain barrier or locally produced by B cells or mast cells contribute to this process. Autoantigens activate the complement cascade, resulting in the formation of the membrane-attack complex and subsequent lysis of the target structure. CD8+ cells are capable of attacking the axon and oligodendrocytes directly. The combination of toxic signals and the disturbed axon-glia interaction pave the way for axonal degeneration. The up-regulation of Ca2+ channels and the increased Ca2+ influx might perpetuate this process. High-frequency signaling of neurons results in axonal degeneration, especially upon exposure to nitric oxide. The loss of signaling activity and trophic support might contribute to axonal degeneration in connected neurons as well. [10]

MS is an autoimmune disease[11] that has a peak onset between ages 20 and 40 years, however it may also develop in children and in addition has been reported in individuals aged above 60 years. MS affected women approximately twice as often as men[12]. The prevalence of MS varies considerably around the world. Kurtzke[13] classified regions of the world according to prevalence that is highest in

northern Europe, southern Australia, and in the middle part of North America. The prevalence in the Italian population shows different rates throughout the country. In particular it was observed a low rate of 53 cases per 100.000 inhabitants in Central and South Continental Italy and a high rate of 81 cases per 100.000 in the northern regions[14]. Central Sardinia prevalence is different from that observed in Continental Italy, with a peak prevalence amongst individuals of 200 per 100.000, a rate that is among the highest in the world[15].

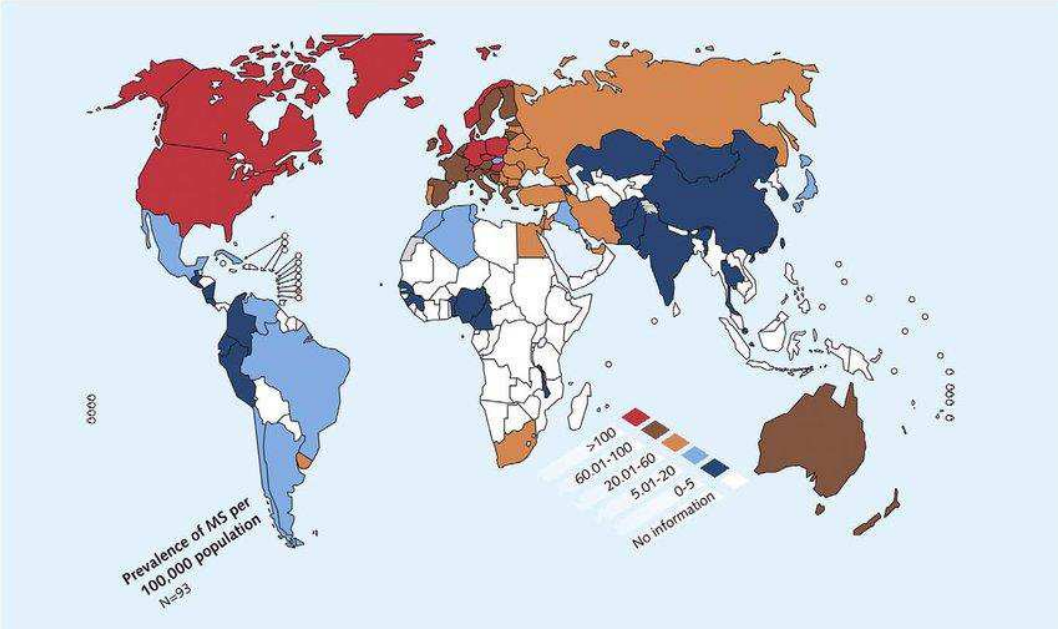


FIGURE 2: The geography of multiple sclerosis: prevalence per 100,000 population.

MS is believed to develop as a result of specific interactions between environmental factors, genetic susceptibility and the development of a pathologic immune-mediated response leading to focal myelin destruction, axonal loss, and focal inflammatory infiltrates. In this scenario are also important the epigenetic modifications, that could play a role in the development and progression of the disease[6]. The prevailing theory is that the disease is triggered by environmental factors but only in those individuals with complex genetic risk profiles.

The pathophysiology of MS is not fully understood yet. Investigators and clinicians who have studied MS agree that the immune system plays a critical role in the

development of lesions, especially during the acute early phases of the disease characterized by relapses. The main pathologic hallmark of MS is the demyelinated plaque, which has specific histological and immunocytological characteristics depending on the activity of the disease. Histologically an MS plaque is characterized by marked predominance of CD8+ T cells and a relative lack of CD4+ T cell. The most recent hypothesis is that CD8+ T cells, γδ cells, natural killer cells, and local antigen presenting cells pass over the blood brain barrier under undetermined circumstances, enter the CNS and cause an immune attack resulting in the inflammatory lesions in CNS. Another important immunopathological feature is continuous synthesis of immunoglobulins in cerebrospinal fluid (CSF), in fact CSF IgG remain one of the most predictive immunological test for the diagnosis of MS[9].

2.1 Genetic factors

Evidence that the genetic factors have a substantial effect on susceptibility to multiple sclerosis is unequivocal. The concordance rate of 31 percent among monozygotic twins is approximately six times the rate among dizygotic twins (5 percent). The absolute risk of the disease in a first-degree relative of a patient with multiple sclerosis is less than 5 percent; however, the risk in such relatives is 20 to 40 times the risk in the general population. The MHC region on chromosome 6p21 is the most important genomic area, well-known to be related to MS[16, 17], but many other regions in other chromosomes are associated to this disabling disease. A genetic map of Multiple Sclerosis (as of 2014) by the International Multiple Sclerosis Consortium is shown in figure 3.

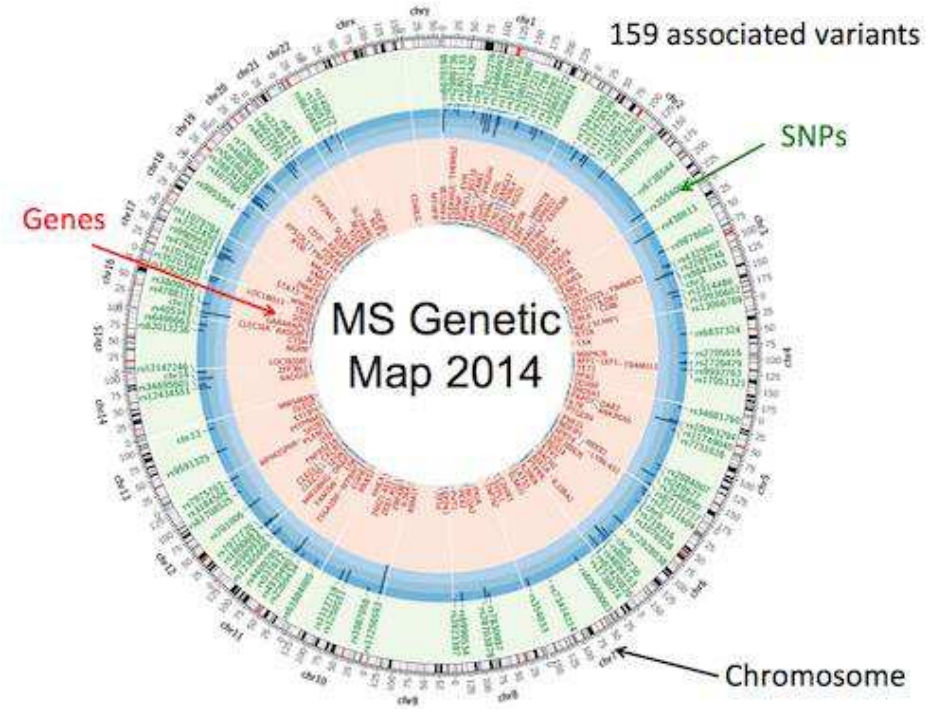


FIGURE:3 MS Genetic Map as of 2014, IMSGC[18].

MS is a multigenic disease involving hundreds of genes and each gene contributes a fraction of the risk factor Also the severity and course of multiple sclerosis is

influenced by genetic factors. Epidemiologic evidence to support this premise comes from studies examining the rate of concordance for measures that describe and quantified variations in the course of disease, including the age at onset, the proportion of patients in whom the disease progresses, and the extent of disability over time. Variants of the interleukin-1β-receptor and interleukin-1-receptor antagonist genes, immunoglobulin Fc receptor genes, and apolipoprotein E gene have been associated with the course of the disease[19].

2.2 Environmental factors

The environment also exerts a significant influence on MS susceptibility[20]. A wealth of studies strongly supports vitamin D deficiency as a key factor for MS. The prevalence of MS correlates with latitude and UV radiation and both vitamin D intake and low vitamin D levels are inversely associated with risk of MS. Possibly the strongest evidence for a role for vitamin D is the association of 2 genes involved in vitamin D biology with MS. As in other immune cell types, vitamin D influences development and functionality of B cells. This pleiotropic hormone plays an important role in B-cell homeostasis and function by decreasing cell proliferation, inducing apoptosis, and inhibiting plasma cell differentiation.

Experimental infection of laboratory animals with various viruses induces demyelination in the CNS. The most studied viral animal model of MS is the disease induced by Theiler's murine encephalomyelitis virus (TMEV), a mouse enteric pathogen that belongs to the single-stranded RNA picornaviruses[21]. The disease model is chronic-progressive in susceptible mice, a striking contrast to the much-used autoimmune EAE model. Two salient features make it the best-suited model for studying MS. There is evidence of an immune response to virally infected cells as well as autoimmune response triggered by viral infection in the CNS, both of which are potentially similar to MS. Epstein-Barr virus (EBV), human herpes virus 6 (HHV-6), varicella zoster virus (VZV), and Chlamydia pneumonia are some of the proposed infectious agents in humans implicated in MS[22].

2.3 Symptoms and current diagnosis of MS

MS has different primary symptoms caused by the loss of myelin and also secondary symptoms triggered by primary MS symptoms. Early symptoms of MS are widely believed to result from axonal demyelination, which leads to the slowing or blockade of conduction.

Most common **primary symptoms** include numbness and weakness in arms, leg and face; chronic fatigue and depression; pain at different levels; vision problems like blurred vision, altered depth perception or even vision loss; disturbed coordination like ataxia (lack of coordination) or tremor (involuntary movement of an arm or leg); bladder and bowel problems and cognitive impairments in memory, mental flexibility, attention and information processing speed. The regression of symptoms has been attributed to the resolution of inflammatory oedema and to partial remyelination. Irreversible axonal injury, gliotic scarring, and exhaustion of the oligodendrocyte progenitor pool may result from repeated episodes of disease activity and lead to progressive loss of neurologic function. Axonal injury may occur not only in the late phases of multiple sclerosis but also after early episodes of inflammatory demyelination.

Secondary symptoms triggered by the primary ones include continuous urinary tract infections due to primary bladder problems or muscle deterioration due to loss of movement in an arm or leg. Since MS does not considerably reduce life expectancy, patients-which are mostly young adults-have to continue their daily lives suffering from these symptoms. Since MS-specific immunoassays are not available yet, the disease diagnosis is mainly based on the clinical history and laboratory investigations proving disease episodes that have affected more than one part of the CNS in more than one occasion and at least one month apart. Therefore, the diagnostic process is significantly longer than many other diseases, taking up to 5 years. Another difficulty concerning the diagnosis of MS is that most of the related clinical features are not MS-specific. Although the use of MRI scan

of the brain and spinal cord is the preferred diagnostic test so far enabling the detection of typical multiple lesions in more than 95% of patients, such kind of lesions may be present in people without the clinical symptoms of MS, in people older than 50 years or in patients of other monophasic disorders like acute disseminated encephalomyelitis (ADEM). Since MRI results by itself are usually not sufficient, additional criteria are used in the diagnosis of MS. One of them is the examination of the CSF by lumbar puncture, which helps to sort out other disease possibilities like chronic infection and vasculitis mimicking an MS appearance. In a CSF test, the most informative analysis is the qualitative assessment of the IgG pattern by using isoelectric focusing with immunoblotting and the comparison with the patient's IgG level in blood. However, neither this test is MS-specific, since elevated IgG levels in CSF are also detected in patients with progressive spinal cord disorders caused by retroviral infections[23, 24].

2.4 Types and treatments of MS

There are several types of MS, differing in their clinical patterns of activity (Figure 4). Relapsing-remitting MS (RRMS) is the most common form of the disease, where symptoms appear for several days to weeks, after which they usually resolve spontaneously. After tissue damage accumulates over many years, patients often enter the secondary progressive stage of MS (SPMS), where pre-existing neurologic deficits gradually worsen over time. Relapses can be seen during the early stages of SPMS but are uncommon as the disease further progresses.

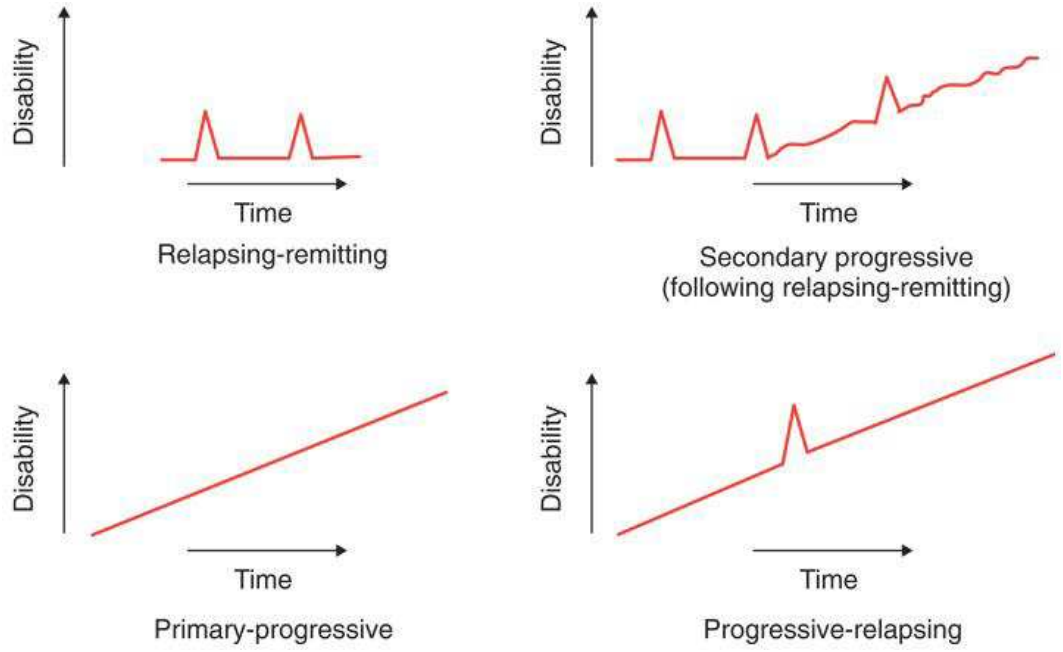


FIGURE 4: Graphical representation of the MS progression types.

About 15% of patients have gradually worsening manifestations from the onset without clinical relapses, which defines primary progressive MS (PPMS). Patients with PPMS tend to be older, have fewer abnormalities on brain MRI, and generally respond less effectively to standard MS therapies[25]. Progressive relapsing MS is defined as gradual neurologic worsening from the onset with subsequent superimposed relapses. Progressive relapsing MS (and possibly a proportion of

PPMS) is suspected to represent a variant of SPMS, where the initial relapses were unrecognized, forgotten, or clinically silent[26].

The course of MS in an individual patient is largely unpredictable. Patients who have a so called clinically isolated syndrome as their first event have a greater risk of both recurrent events and disability within a decade if changes are seen in clinically asymptomatic regions on MRI of the brain. The presence of oligoclonal bands in cerebrospinal fluid slightly increases the risk of recurrent disease[27]. Ten percent of patients do well for more than 20 years and are thus considered to have benign multiple sclerosis. Women and patients with predominantly sensory symptoms and optic neuritis have more favourable prognosis. Life expectancy may be shortened slightly, in rare cases, patients with fulminant disease die within months after the onset of MS. Suicide remains a risk, even for young patients with mild symptoms[28].

2.5 Multiple sclerosis in Sardinian population

Repeated epidemiological assessments of MS in the Mediterranean island of Sardinia have demonstrated how incidence of the disease is among the highest in the world, with an incidence rate of 4.2 and prevalence of 152 per 100,000 inhabitants[15]. The high incidence of the disease on the island likely derives from the particular genetic makeup of Sardinians and from the founder effect, possibly underlying transmission of MS-susceptible genetic material originating from few common ancestors[29]. A recent report has demonstrated a marked increase in MS incidence on the island over the last three decades. In addition to the temporal trend of MS, an anticipation of age at onset has been reported in Sardinia over the past 50 years. The latter phenomena may have been influenced not only by stochastic and epigenetic events but also by environmental variants interacting with the individual genetic makeup[30].

Considering these facts, the Sardinian population has been counted as a special population in the context of understanding the relative contribution of environmental and genetic factors for the development of MS. This homogeneous population represents an ideal dataset[31-33] for studies like the one presented here with the aim of identifying novel causal pathways that could help develop new therapeutic strategies and better understand the pathogenic aspects of the disease, as they may shape the development and progression of the disease itself[15, 34].

3. What is "causality"? A brief history

Causa latet: vis est notissima.

- Ovid

3.1 The "causal" problem

The very central aim of most of the studies in health, social science and economic research, is to learn about causal explanations of the data, i.e. to unfold cause-effect relationships among variables and/or events.

So, causal analysis goes one step further of standard statistical analysis, which consists in assessing parameters of a distribution from samples drawn of that distribution, inferring with the help of such parameters associations among variables, estimate beliefs or probabilities of past and future events. Causal analysis aims in fact to infer not only beliefs or probabilities under static conditions, but also the dynamics of beliefs under changing conditions, as changes induced by treatments or external interventions[35].

Understanding the world and not limiting ourselves to describe it is what causality is, in the end, about.

The problem of discriminating, with a certain amount of confidence, between correlation and causation is therefore key in medical research as in many other biological and social sciences: what is the cause of a specific disease? How does smoking affects someone's health? What could be the effect of a certain kind of treatment? Which type of interventions could be helpful in improving social issues like micro-criminality? Which educational programs would most benefit children in schools? All these kinds of questions have always been of primary interest for

researchers of all fields, and all of them implies some sort of causality notion and needs to answer questions of cause-effect relationships.

However, how this ambitious task could be achieved has been for decades a matter of lively discussion.

Especially in biostatistics and epidemiology though, cautiousness has been traditionally the main approach to such problematic topic, and researchers used to very rarely even dare to use or pronounce the word “causal”, stressing the mathematical and statistical requirements to be able to give confident and secure conclusions, something that ideally was only achieved by controlled experimental studies[36].

In addition, bringing more vagueness to the field, it must be said that there did not even exist any rigorous translation between the language of causality and the language of probability distributions[37]. This lack of a specific semantic and mathematical language to cast and derive causal questions and answers lasted for several decades ‘till recent years and helped (let’s say caused!) to delay a more decisive and formal approach to the issue.

Another reason why there’s been much more talks about associations rather than causes comes indeed from philosophy. In fact, even philosophers of science couldn’t (can’t) agree on what properly constitutes a “cause”.

Narrowing the view only to western philosophy, starting from Plato and more substantially Aristotle, with his four causes types (material cause, efficient cause, final cause and formal cause), going through the Stoic principle of “universal causation” straight to modern philosophers, in particular Hume with his formalization of the three factors characterizing causal relations (*contiguity* of cause and effect, *priority* in time of cause to effect and *necessary* connection between cause and effect), and their everlasting diatribe between rationalists and the empiricists, almost everyone has tried to disentangle the problematic issue of “causality” giving his own view and opinion about the - so hot - topic[38].

All these are just some of the aspects that contributed to the struggling history of causality through the ages.

3.2 Fisher and the Design of Experiments

Coming back to contemporary life sciences and statistics though, one of the first and most important effort to tackle the problem of which are the essential elements to achieve reasonably secure causal conclusions was done in Harpenden, by Fisher[39, 40] (and later by Yale[41, 42] and others). There, at Rothamsted Experimental Station, one of the oldest agricultural research institutions in the world, were laid the foundations of the theory of experimental design, developed mostly in the context of agronomy studies. During those years, among others were highlighted some fundamental principles that would have been key for the consequent development of the topic: clear definition of experimental units (which in most of Rothamsted experiments were plots of soil) and treatments (as before, a specific fertilizer), assumptions of unit-treatment additivity, and in particular what later would be defined as conditioning on features prior to treatment allocation and marginalization of features between treatment and final outcome (the absence of what is nowadays defined *noncompliance*: any intervention between treatment allocation and response should either be independent of the treatment or reasonably defined as intrinsic part of it)[43].

Briefly, that's a simplistic example of what Fisher and the other researchers were studying in those days at Rothamsted: aiming to determine whether the addition of a nitrogen-based fertilizer could cause an increase in the seed yield of a particular variety of wheat, a randomized experiment was designed in which a field was divided into plots of soil; to each of them was then randomly applied or not applied the fertilizer (the treatment variable). No further manipulations were then involved until harvest day, when the seed that was harvested from each plot was weighted accurately.

By randomizing the treatment allocation, a sampling distribution is generated that allows to calculate the probability of observing a given result by chance if, in reality, there is no effect of the treatment.

The great and novel insight given by Fisher was that the process of randomization could ensure, up to a probability that could be calculated from the sampling distribution produced by the randomization, that no uncontrolled common cause of both the treatment and the response variables could produce a spurious association[44]. "*Randomisation properly carried out […] ensures that the estimates of error will take proper care of all such causes of different growth rates, and relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed*"[40].

It's however unnecessary to say that Fisher's work did not put an end to controversy and further debates; on the contrary, it was the spark that ignited a heated and lively discussion.

3.3 Development of Causal Inference Theory

From then, many renown authors gave their contributions to the development of the topic, facing also the problem of causality in non-experimental studies.

Cochran[45], Cox[46] and Cox & McCullagh[47] suggested methods involving the use of supplementary variables to improve efficiency of estimator and the implementation of instrumental variables to make identifiable a causal effect of interest.

Cochran[48] gave also a huge contribution exploring many aspects of the analysis of observational studies (of which more will be said in the next chapter) and reported Fisher's reply to a question that he had asked him about how to make observational studies more likely to yield causal answers: "Make your theories elaborate", a quote that would have become ever-present in causality debates from then on.

In 1965, Bradford Hill[49] contributed to develop a theory of causal inference in observational studies, proposing a set of guidelines:

1) *Strength of association*: a strong association is more likely to have a causal component than a modest association
2) *Consistency*: a relationship is observed repeatedly
3) *Specificity*: a factor influences specifically a particular outcome or population
4) *Temporality*: the factor must precede the outcome it is assumed to affect
5) *Biological gradient*: the outcome increases monotonically with increasing dose of exposure or according to a function predicted by a substantive theory
6) *Plausibility*: the observed association can be plausibly explained by substantive matter (e.g. biological) explanations

7) *Coherence*: a causal conclusion should not fundamentally contradict present substantive knowledge
8) *Experiment*: causation is more likely if evidence is based on randomized experiments
9) *Analogy*: for analogous exposures and outcomes an effect has already been shown

Satisfaction of some or all of them would strengthen the case for causality inferred from observational studies. Although he did not provide a specific definition of "causal", the relevance of this work has been huge in the epidemiological world.

Rubin[50], developed the causal notation of *counterfactuals* or *potential outcomes*, already introduced by Neyman in 1923[51]. This framework had the merit to finally give a formalization to some intuitive approach to causality that were commonly used in the field. Specifically, let Y_x denote the outcome for a random subject in the study population, if the subject would hypothetically receive exposure level *x*. Depending on what exposure level the subject actually receive, Y_x may or may not be realized, and is referred therefore to as a *potential outcome*.

Let X and Y denote for a given subject the observed exposure and outcome; if the subject is exposed to level $X = x'$ then the potential outcome $Y_{x'}$ is assumed to be equal to the observed factual outcome *Y*. The link between potential and factual outcomes is usually referred to as "consistency assumption", and is formally expressed as

$$X = x' \rightarrow Y_{x'} = Y$$

Thus, for a subject exposed to $X = x'$ all potential outcomes except $Y_{x'}$ are unobserved, or *counterfactual*, echoing the fact that all the unobserved potential outcomes correspond to hypothetical scenarios that did not happen, that are "contrary to fact", which makes those subject-specific causal effects in effects in general not identifiable. What can be done in such framework, though, is measuring an aggregate impact of the exposure over the whole study population, a *population causal effect*. In fact, the potential outcome Y_x may vary across

different subjects, and can be treated therefore as a random variable following a probability distribution Pr(Y_x).

Dawid[52], criticized potential outcomes for unnecessarily suggesting that Y(*0*) and Y(*1*) (example in the case of a binary treatment/exposure) simultaneously exist and therefore have a well-defined joint distribution also arguing that the retrospective nature of potential outcomes is unscientific and unnecessary and therefore proposed a decision-theoretic causal inference framework instead. Though deeply different in their philosophical fundamentals (or lack of), the methods appear to practically suggest the same analysis, though the decision-theoretic one allows to relax more the strong assumptions behind potential outcomes.

Robins[53, 54], in effect explored notions of causality in a clinical trial and epidemiological setting.

Rosenbaum[55] has given a searching discussion of the conceptual and methodological issues involved in the analysis of observational studies.

Pearl[35], with all of these authors, finally succeeded in these last decades to transform cause-effect relationships into objects that can be manipulated mathematically, formalizing fundamental concepts like confounding, setting up various frameworks for causal inference from both experimental and observational studies[56, 57].

3.4 Independence and conditional independence

Talking about formalities, it's time to briefly introduce some fundamental concepts in the theory of statistical inference: *independence* and *conditional independence*.

Following Dawid's work[58], conditional independence offers a new language for the expression of statistical concepts and a framework for their study, achieving a unification of many area of statistics which appear, at first sight, to be entirely unrelated. It involves two of the most basic concepts in statistics: independence and conditional probability.

Independence: Let X and Y be random variables. We denote by *p(x,y)* the joint density of (*X,Y*), by *p(x)* the marginal density of *X* and by *p(x|y)* the conditional density of *X* given *Y* = *y*. X⫫Y denote that *X* and *Y* are independent, so in term of density we have

$$p(x,y) = p(x)p(y)$$

$$p(x|y) = p(x)$$

and expressing a factorization of p(*x,y*) we have

$$p(x,y) = a(x)\,b(y)$$

$$p(x|y) = a(x)$$

All the above formulations of the property of X⫫Y are mathematically equivalent.

Conditional Independence: Introducing a further variable *Z*, we use the notation *X⫫Y|Z* denoting that *X* and *Y* are probabilistically independent in their joint distribution given *Z* = *z*, for any observable value z of Z. Again, we can deduce the following equivalent expressions of this property:

$$p(x,y|z) = p(x|z)p(y|z) \qquad \text{(a)}$$

$$p(x|y,z) = p(x|z) \qquad \text{(a2)}$$

$$p(x,y|z) = a(x,z)\,b(y,z) \qquad \text{(b)}$$

$$p(x|y,z) = a(x,z) \qquad \text{(b2)}$$

The content of $X \perp\!\!\!\perp Y|Z$ is probably best captured by (b2), which clarifies that the conditional distribution of *X*, given *Y* and Z, is in fact completely determined by the value of *Z* alone, so the value of *Y* provides no further useful information to predict the value of X. It must be noted that the conditional independence relation $X \perp\!\!\!\perp Y|Z$ does not necessarily imply the marginal independence $X \perp\!\!\!\perp Y$. Situations in which the two independence conditions are not both respected are common, meaning that two independent variables in the population, *X* and *Y*, could become dependent when we observe (condition on) the precise value of *Z*.

Bearing these principles clear in mind can help assessing many problems in the framework of causal inference giving a handful approach in formalizing most of the causal questions and relationships among the variables under study.

3.5 Causal Diagrams

If it's handful approaches that we're looking for, then probably nothing can best graphical representations. Causal diagrams[59], deriving from path analysis and structural equations modeling, have been informally used for long time in various fields of research, and in past years underwent a proper formal development also in epidemiologic research, giving birth to the theory of directed acyclic graphs (DAGs). Graphs can provide a useful starting point for identifying variables that must be measured and controlled to obtain unconfounded effect estimates, to design and program studies and projects, and they also provide a method for critical evaluation of traditional epidemiologic criteria for confounding.

When investigating any causal relationship, it is necessary to start from some set of causal assumptions (analysis model) pertaining to causation, measurement, selection and probability distributions. Common statistical models require the introduction of many parametric assumptions, often untestable or not easy to test for; graphical models (also called influence diagrams, relevance diagrams, or causal networks), instead, allows to avoid those strong parametric assumptions incorporating assumptions about the web of causation not captured by conventional ones.

We will here enounce some of the basic steps to follow for the creation of a causal graph along with the fundamental definitions and principles used in graph language.

We can create a causal diagram abstracting the causal assumptions implied in a description of the hypothetical relationships existing among the studied variables. The points on the graph representing the variables are called *nodes* or *vertices* (A, B, C, D, E in the example below, Figure 5). Any line or arrow connecting two variables in the graph is called an *arc* or an *edge*.

Two variables in a graph are *adjacent* if they are directly connected by an arc (in the example A and C are adjacent but A and D are not). Single-headed arrows represent direct links from causes to effects, not mediated by other variables. For example, in our graph the direct effect corresponds to the arrow linking A to C, that is a direct effect not mediated by other variables. In the same way, the absence of directed arrows means absence of direct effects of a variable on another.

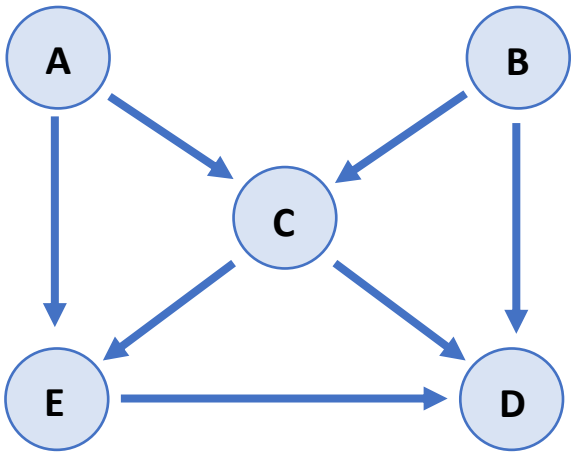


FIGURE 5: Example of a directed causal diagram.

A *trail* is a sequence of distinct vertices forming a path in the graph. A *path* through the graph is any unbroken route traced out along or against arrows or lines connecting adjacent nodes.

A *directed path* from one node to another in the graph is one that can be traced through a sequence of single-headed arrows, always entering an arrow through the tail and leaving through the head; this path is also called *causal path* in causal graph. In our example the path A-C-D is directed, instead E-C-D is not.

A variable is an *ancestor* or cause of another variable if there is a directed path of arrows leading out of the first into the second; in this case the second variable is

defined a *descendent* of the first or affected by it. In the example A, B and C are ancestors of E and D and E and D are therefore descendants of A, B and C.

A variable *intercepts* a path if it is in the path (but not at the ends). Variables that intercept directed paths are *intermediates* or *mediators* on the pathway.

A bidirectional (two-headed) arrow, connecting two variables in a graph, is often used to indicate that the two variables share an ancestor (have a common cause), but the ancestor and their interrelations are not shown in the graph. This common ancestor may represent more than one variable. Non-directional dashed line represents relations whose source is not specified by the graph.

A node *v* of a trail π is a *collider* in π if it is an internal node of π, and the arrows of π meet head-to-head at *v*. Let S be a subset of the graph nodes: a trail π is said to be *blocked* by S if it contains a node *c* such that either

- $c \in S$ and *c* is not a collider in π

or

- *c* and all its descendants are not in S, and c is a collider in π.

A trail that is not blocked by S is said to be *active*, two subset A and B are said to be *d-separated* by S if all trails from A to B are blocked by S.

A graph is *acyclic* if no directed path in the graph forms a closed loop, so it is not possible starting from a node and following the arrow, back to that node (it no contains cycle). A *directed acyclic graph* (DAG) represents a complete causal structure, in that all sources of causal dependence are explained by causal link and therefore all edges between pairs of nodes have a direction (arrows)[60].

Although causal diagrams provide results similar or equivalent to those obtainable using classical counterfactuals causal models, they seem to have a much more intuitive appeal, and can be very useful for identification and control of confounding or measurement processes. Nonetheless, even graphical approaches do not free us from the heavy burden of basing ourselves on solid assumptions, non-derivable from observational data alone.

3.6 Assessing causality in observational studies

Although the randomised experiment is considered the "gold standard" for making and testing causal hypothesis, it can't be properly applied to many (perhaps most) research questions asked by biologists. The strength of the randomized experiment is in the fact that we do not need to physically control (or even care) for other causally relevant variables in order to reduce the possibility that the observed association is due to some unmeasured common cause in our sample[61].

Anyway, this kind of studies in some cases present insurmountable challenges to the researcher: treatments only targeting risk factors of interest may be difficult (or even impossible) to find, many types of treatments can't be randomly allocated due to ethical or practical reasons (think about a study on the effects of alcohol consumption on the risk of developing cancer), they are generally quite expensive and time-consuming (especially studies that involve long-termed follow-up of subjects). In all these scenarios, alternative approaches to assess causal relationships must be adopted, also relying on observational data.

As defined by Cochran[48], in an observational study "the objective is to elucidate cause-and-effect relationships, or at least to investigate the relationships between one set of specified variables x_i and a second set y_i in a way that suggests or appraises hypotheses about causation".

For the most part, indeed, an observational study is a study of the associations between two sets of variables. Attempts to interpret these associations as causal or non-causal must rely heavily on information not supplied by the study, though some information may come from previous results.

There are, obviously, difficulties involved. The most familiar difficulty is that the treated and control groups (exposed and unexposed subjects) may not be directly comparable, since treatments were not randomly assigned to experimental units.

Even after adjusting or controlling for observed covariates (i.e., measured characteristics prior to treatment allocation), estimates of treatment effects can still be biased by imbalances in unobserved covariates[48, 62, 63].

Replication of the results is always strongly recommended before making decisive conclusion on causality, but observational studies with different strengths and limitations may or may not corroborate one another.

Even within a single observational study, though, it is often possible, to provide some assessment of the evidence about the causal effects of a treatment.

To this extent, Rosenbaum[64] for example suggested a couple of applicable methods which included covariance adjustment (subclassification, matched sampling or related methods used to adjust for observed covariates), checking the consistency of the assumption of *strongly ignorable treatment assignment* (in the absence of such consistency, we cannot safely rely on standard methods of adjustment to produce appropriate estimates of treatment effects), examining the sensitivity of estimates to assumptions about unobserved covariates (if estimates are relatively insensitive to plausible variations in assumptions about unobserved covariates, then a causal interpretation is more defensible).

Some other useful suggestion can be found in Cochran's work[48]. First of all, the author recalls the Fisher's hint "Make your theories elaborate", meaning that when constructing a causal hypothesis, as many different consequences of its truth as possible should be envisaged, and observational studies to discover whether each of these consequences is found to hold be planned in consequence.

In addition, since most of the studies are often conducted on restricted populations, repetition of the study plan in different environments by different workers can be valuable, especially in understanding whether results can be extended to a broader target population. Since similar studies may be subject to the same biases, an approach with a different plan that escapes some of these biases could be highly useful.

In conclusion, evidence on a question that has to be decided mainly from observational studies will usually consist of a heterogeneous collection of results of varying quality, each bearing on some consequence of the causal hypothesis. Whenever results are contradictory, reaching a verdict surely demands much skill. Assessing causal relationship using observational data is therefore no trivial task, and there's always need for more robust approaches. Mendelian randomization, the one we chose to conduct our study, is one of these[65].

3.7 Mendelian Randomization

As already mentioned, assessing causality in observational studies forces to face many different problematic issues, relying on untestable and often implausible assumptions such the absence of unknown confounders and of reverse causation (the outcome Y itself causes changes in the exposure level X and not vice-versa). Confounding is often due to the existence of another cause of Y that is also associated with X and controlling for confounding is problematic, as often they are unknown or unmeasured making it impossible to account for them in the analysis.

Mendelian Randomization (MR) framework has been properly proposed to address this kind of problems typical in classic epidemiology.

MR has been defined as "instrumental variable analysis using genetic instruments"[66], and therefore it's useful to start presenting what is an instrumental variable (IV).

An IV is a measurable quantity which happens to be associated with the exposure of interest, and not associated with any confounding factor nor with the outcome except via the hypothesized causal pathway passing through the exposure itself. These conditions constitute the fundamental assumptions that must be respected by a variable to be considered a valid instrument. Being able to identify reliable IV can give rise to natural experiment conditions, from which can be confidently inferred causal relationships.

In MR, genetic variants, simplistically section of the genetic code that differs between individuals, are used in causal inference studies as Instrumental Variables[67]. That's because genetic variants have some particular characteristics that make them fits particularly well with the previous mentioned assumptions: genetic variants are in fact generally randomly distributed (alleles randomly inherited at meiosis) in the population (except in presence of non-random mating and selection effects, which are very rare situations)[68] independently of any other

variables, so that subgroups of subjects bringing same variants do not differ systematically with respect to any of these. This makes the MR framework analogous to a RCT[69], in particular inferring a causal effect of the exposure on the outcome recalls the estimation of an intention-to-treat effect in a RCT. Moreover, the genetic code of each individual is already "set up" at birth, denying every possibility of a variant to be caused by some variable measured in mature age and protecting therefore from reverse-causation possibilities.

For the MR framework to work, it is crucial though for the necessary assumptions of instruments validity to be respected:

- the variant is associated with the exposure

- the variant is not associated with any confounder of the exposure-outcome association

- the variant does not affect the outcome, except potentially via its association with the exposure

As in all observational scenarios, almost always these assumptions are untestable, and researchers have to rely most on basic biological knowledge and previous findings.

In terms of random variables, assuming that we have an outcome Y, function of a measured exposure X and unmeasured confounder U, being G our set of genetic variants (one or more), these three assumptions can be translated into the language of conditional independencies as follows:

- $G \not\perp\!\!\!\perp X$ (G is not independent of X)
- $G \perp\!\!\!\perp U$ (G is independent of U)
- $G \perp\!\!\!\perp Y \mid X, U$ (G is independent of Y conditional on X and U)

which corresponds to the following graph[70, 71] (Figure 6):

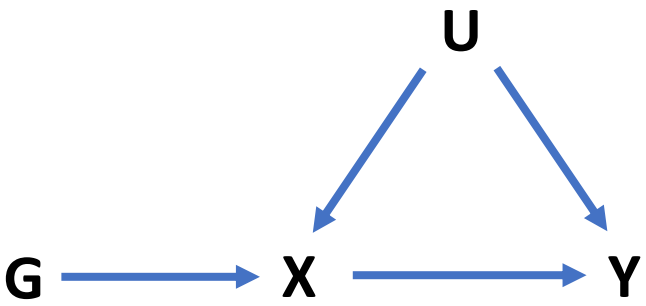


FIGURE 6: Directed Acyclic Graph illustrating Instrumental Variable assumptions.

The first assumption guarantees the different subgroups defined by the variant to have different levels of exposure, implying systemic differences between them.

A variant which is not strongly associated with the exposure is referred to as a *weak instrument*, which differs from an invalid one to the extent that weak instruments can be made stronger collecting more data and still give valid conclusions (even if maybe with low power in detecting true causal effects).

The second assumption reassure that all the other variables in play (observed and unobserved) are equally distributed among all the subgroups, making comparisons between them reliable.

The last assumption, in the end, means that that are not any other pathways linking the variant to the outcome other than through the exposure.

Violations of one or more of the assumptions would lead to invalid instruments and to biased estimates of causal effects; nonetheless reasons that could lead to violate them are different, ranging from biological mechanisms, like pleiotropy and canalization, to linkage disequilibrium, to population effects like stratification.

We will briefly describe some of the most common:

Pleiotropy: gene or genetic variants can have more than one independent phenotypic effect. When a genetic variant is associate with multiple different risk-

factors for the outcome, then either IV assumption 2 or 3 is violated and the variant is therefore invalid

Linkage disequilibrium: ideally all loci on the genome exhibit a complete independence in the population (i.e. are in linkage equilibrium). Variants which are physically near to each other on the same chromosome though could be inherited together, showing correlated distributions. If some of these correlated variants are associated with competing risk factors, IV assumptions are violated in a similar way as the pleiotropy case.

Population stratification: if a population under study can be divided into distinct subpopulations, for example typically when there are subgroups of different ethnic origin, the frequency of the genetic variants could differ substantially between the groups. This could lead to misinterpret an association caused from these differences attributing it to the genetic variant.

Methods to face all those problematic limitations have been developed in the Mendelian Randomization framework, using single as well as a plurality of genetic variants, some of which like summary statistics methods used in this study will be treated more specifically in the methods section.

4. Materials and Methods

Take a method and try it. If it fails, admit it frankly, and try another. But by all means, try something.

- Franklin D. Roosevelt

4.1 Dataset

4.1.1 Sample collection and genotyping

MS patients were ascertained through the case register established in 1995 in the province of Nuoro, Sardinia. All the cases were diagnosed according to Poser criteria[24]. Twenty extended pedigrees containing from three to eleven MS patients were selected for the analysis. The overall structure of the selected twenty pedigrees includes 98 cases and 838 unaffected relatives. Of these 936 subjects, 268 (212 controls and 56 case) have genotype data; 211 (142 controls and 69 cases) have only protein data, while 131 (90 controls and 41 cases) had both protein and genotyping data.

Genotype data were obtained using Immunochip from a previous study[72], in which the quality control-filtered dataset included 131.497 SNPs. A pruned set of 19.121 independent SNP ($r2 < 0.20$ within a windows size of 100 Kb) was obtained using PLINK's indep-pairwise command and used in the MR analysis, guaranteeing no LD presence among the SNPs brought to the final analysis.

4.1.2 Measuring concentration of plasma protein

Proteome in plasma has started to be studied in a high throughput manner, thanks to the technological development of the Human Protein Atlas (HPA) project which allowed to produce a large resource of polyclonal antibodies (mAbs)[73] (www.proteinatlas.org).

The availability of an antibody array format called Suspension Bead Array (SBA), consisting of HPA antibodies immobilized onto microspheres in suspension, allows to perform plasma analysis on up to 384 proteins[74, 75].

Antibody-based procedure based on the suspension bead array technology enables a unique analysis of body fluids with a high multiplexing degree. Through a setup that profiles proteins via direct labelling of whole and unfractionated samples with biotin, any given target can be addressed with an antibody. For the technology of choice, antibodies are coupled to colour-coded magnetic beads to create antibody suspension bead arrays of a desired composition. With direct access to more than 22000 antibodies, assays can be performed in a manner yet unmet by any other affinity-based procedure to profile proteins in body fluids. Instant data acquisition from the flow cytometer, information can be directly uploaded for further data analysis. Measurements are performed with the Luminex LX200 (temperature optimization) or the Flexmap 3D instrument (biomarker discovery). To determine relative signal intensities from the binding of antibodies to their target antigens, Median Fluorescent Intensities (MFI) are displayed when counting at least 50 events per bead ID. The obtained data is processed using Probabilistic Quotient Normalization (PQN)[76] over the entire data set to account for possible differences in sample dilution.

A list of potentially MS related proteins had been put together by clinical collaborators and were available for antibody selection. For temperature optimization, 78 mAbs antibodies targeting 73 different proteins were utilized; 68 antibodies were selected from suggestions by the clinical collaborators and 5

control antibodies were targeting abundant proteins in plasma and CerebroSpinal Fluid (CSF) and are TTHY, FIBB, APOA4, CFB and C3. One nonspecific anti-rabbit IgG (Jackson Immunotech) and one albumin binding protein (HisABP) were also employed as negative and positive controls, respectively. For the biomarker discovery study, 379 mAbs were included. 152 of these antibodies were selected from suggestions by the clinical collaborators and 176 antibodies from proteomic, mRNA and cDNA expression data found in the literature.

Moreover, since new antibodies produced within the HPA project are routinely coupled to a 384-plex suspension bead array to profile plasma samples in a multi-disease cohort, 56 antibodies that were identified having different relative protein levels in a MS cohort as compared to healthy control samples were among the 377 selected mAbs. Four HisABP (one per bead coupling plate), one anti-rIgG and several control antibodies with high abundance in CSF and plasma were also included. In the resultant list of 377 mAbs, 337 were targeting unique proteins, all antibodies with a concentration >0.04 mg/ml.

In our opinion, this selection, directly made on the basis of expert clinicians' opinion and literature review on the subject, constitutes an optimal starting point for our study, and can be reasonably considered a plausible set on which our causal investigation can be conducted, hopefully leading to, after the implementation of a rigorous and specific statistical analysis, the detection of interesting suggestions of new (or confirmation of already known) putative causes and targets of the disease.

From now on, for clarity, readability and uniformity with the original dataset, we will refer to each specific antibody targeted product, the real object of sampling and analysis, as a single "protein".

4.2 Selection of IVs

The framework of causal inference, more specifically MR that uses genetic variants as IVs, is one of the approaches to tackle the problem of confounding in observational studies. Indeed, the association between the genetic variant and the disease is not subjected to confounding since the genetic variant has been randomized during the meiosis. Under the concept of MR, the genetic variants allow to assess the possible causal effect of X on Y, without requiring any experimental intervention on X[1, 77, 78].

To be a valid IV for the estimation of the causal effect of interest, the genetic variant has to satisfy three core assumption: a) to be associated with X, b) to be independent of Y conditional on X and on the confounders of the relationship between X and Y; this latter condition means that the IV affect Y only via its effect on X and, lastly, c) to be independent of the confounders of the relationship between X and Y. No pleiotropic effect of the IV on Y exists when IV satisfies condition b) and c).

Only the first condition is empirically verifiable, by regressing Y on X, as the last two involve the unobserved and/or unknown confounders, and thus they cannot be tested. In general, when a genetic variant violates these assumptions is considered as invalid and its inclusion in the analysis may lead to biased estimates. A further requirement is that the IVs must be independents. In fact, one way in which confounding could be reintroduced into MR studies is to use IVs that are in strong Linkage Disequilibrium (LD). When a locus under study is in LD with another polymorphic locus confounding will result if both the loci are associated with the outcome of interest. However, using independent genetic variants, which both serve as proxies for the risk factor of interest makes much less plausible that reintroduced confounding explains the association, since it would have to be acting in the same way for these unlinked variants. The use of multiple genetic

variants working through different pathways therefore represents a way to avoid this kind of problem[79].

By combining the estimates of association from multiple variants into a single estimate of the causal effect, an assumption is made that the variants provide independent information. In addition, if two genetic variants are in complete LD, then inclusion of both variants in the model would not lead to additional information. In theory, variants in partial linkage equilibrium could provide additional information on the causal effect, and information on such variants could be included correctly in a summarized analysis, but highly reliable methods to implement that kind of information have not been developed yet. Therefore, variants used in a summarized analysis must be uncorrelated in order to obtain valid statistical inferences analyses.

We performed two MR analysis: 1) the main analysis to assess the causal effect of the protein level on MS, 2) and the bidirectional MR analysis to investigate a possible reverse causation.

All the analyses were performed adding sex as covariate, accounting for the familiar relationship between subjects and choosing a priori significance threshold of $p<5x10^{-4}$ to identify and select significant SNPs from the two models. These significant signals composed different lists of IVs to be used respectively for the main MR analysis and for the bidirectional MR analysis.

4.2.1 Protein ~ SNPs association

For the main MR analysis, we needed a sufficient number of independent SNPs to be used as IVs, significantly associated with the level of each candidate plasma protein. To this aim we fitted linear mixed-effects kinship models between each protein profile and each SNPs from the pruned set of independent Immunochip genetic variants, using coxme R package.

In particular, we used a function which is an implementation of linear mixed effects models able to fit models with random family effects, i.e., using a kinship matrix for the correlation, called "lmekin".

The random effects linear model is:

$$y = X\beta + Zb + \varepsilon$$

$$b \sim N(0, \sigma^2 A(\theta))$$

$$\varepsilon = N(0, \sigma^2)$$

Here β are the fixed and b the random coefficients, and the variance matrix A of the random effects depends on some arbitrary vector of parameters θ. For any fixed value of θ the solution for the remaining parameters is based on a QR decomposition[80].

For known A, this is solved as an augmented least squares problem with

$$y^* = \begin{pmatrix} y \\ 0 \end{pmatrix} \qquad X^* = \begin{pmatrix} X \\ 0 \end{pmatrix} \qquad Z^* = \begin{pmatrix} Z \\ \Delta \end{pmatrix}$$

where $\Delta'\Delta = A^{-1}$. The dummy rows of data have $y = 0$, $X = 0$ and Δ as the predictor variables. With known Δ, this gives the solution to all the other parameters as an ordinary least squares problem, which is solved using a QR decompostion. The Z matrix is often sparse, so the QR computations are done using the Matrix library to take advantage of this. Maximization of $L(\theta)$ with respect to θ is accomplished with the optim() function.

4.2.2 MS ~ SNPs association

For the bidirectional MR analysis to test for the presence of reverse causation, we need sufficient number of independent SNPs, to be used as IVs, significantly associated with MS. This analysis was performed using GWAF R package[81], which allows to fit logistic regression via Generalized Estimation Equation (GEE) and to

test associations between a dichotomous phenotype and each SNP using an independence working correlation matrix, with each family acting as a cluster in the robust variance estimation of the genotype effects.

Liang and Zeger[82] first introduced the GEE approach to treat with correlated data.

Consider a sample of i = 1, . . . , *K* independent multivariate observations $Y_i = (Y_{i1}, . . . , Y_{it}, . . . , Y_{ini})$. Here *i* may represent a cluster with n_i observations.

The expectations $E(Y_{it}) = \mu_{it}$ are related to the *p* dimensional regressor vector x_{it} by the mean-link function g

$$g(\mu_{it}) = x_{it}T\,\beta \qquad (1)$$

Let

$$\mathrm{VAR}(Yit) = \varphi a_{it} \qquad (2)$$

where ϕ is a common scale parameter and $a_{it} = a(\mu_{it})$ is a known variance function. Let $R_i(\alpha)$ be a working correlation matrix completely described by the parameter vector α of length m.

Let

$$Vi = \varphi A_i^{1/2}\, R_i(\alpha) A_i^{1/2} \qquad (3)$$

be the corresponding working covariance matrix of Y_i, where A_i is the diagonal matrix with entries a_{it}. For given estimates $(\hat{\varphi}, \hat{\alpha})$ of (ϕ, α) the estimate $\hat{\beta}$ is the solution of the equation

$$\sum_{i=1}^{k} \frac{\partial \mu_i^T}{\partial \beta}\, V_i^{-1}(Y_i - \mu_i = 0) \qquad (4)$$

Liang and Zeger suggest using consistent moment estimates for ϕ and α. This yields an iterative scheme which switches between estimating β for fixed values of $\hat{\varphi}$ and $\hat{\alpha}$ and estimating (ϕ, α) for fixed values of $\hat{\beta}$. This scheme yields a consistent estimate for β. Moreover, $K^{1/2}(\hat{\beta} - \beta)$ is asymptotically multivariate normally distributed with zero mean and covariance matrix

$$\sum = \lim_{K\to\infty} \sum_0^{-1} \sum_1 \sum_0^{-1} \qquad (5)$$

where

$$\sum_0 = \sum_{i=1}^{k} \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T} \quad ; \quad \sum_1 = \sum_{i=1}^{k} \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} COV(Y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T} \quad (6)$$

Replacing β, ϕ and α by consistent estimates and the covariance matrix COV(Y_i) by $(Y_i-\mu_i)(Y_i-\mu_i)^T$ in (6) yields a so called sandwich estimate $\hat{\Sigma}$ of Σ. The estimate $\hat{\Sigma}$ is a consistent estimate of Σ even if the working correlation matrices $R_i(\alpha)$ are misspecified[83].

4.3 Mendelian Randomization and Bidirectional MR analysis

We use the influence diagram (Figure 7) to graphically represent the general problem of interest, the MR approach and the assumptions. We consider X as the plasma protein level and Y as the binary outcome given by absence or presence of MS. Z_1, Z_2 and Z_j in the graph represent the multiple uncorrelated IVs previously selected (see section Selection of IVs) that, under the principle of MR, allow us to assess the existence of a putative causal effect of X on Y. In the diagram are also reported the unknown and/or unobserved confounders, U, of the relationship between X and Y. The influence diagram takes also into account the problem of pleiotropy that appears when the association between the IV (in the graph represented by Z_1) and Y is not entirely mediated by the studied X (i.e. Z_1 →Y direct arrow).

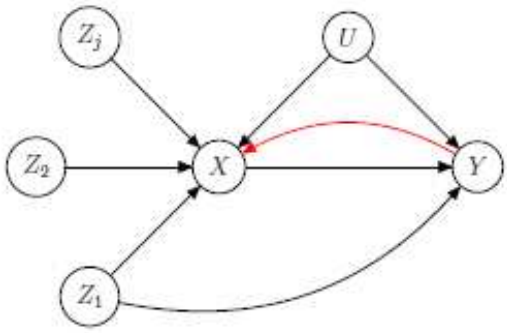


FIGURE 7: **Influence diagram representation of the general problem of interest in this work.** The arrows in the diagram represent putative cause-effect relationships, and the nodes represent random quantities in the problem. The red arrow represents the effect of Y on X (i.e. reverse causation).

For the main MR causal analysis, performed for each candidate proteins, we used different MR methods: Inverse-Variance Weighted (IVW) as primary method for

the main causal analysis, MR Egger regression (MR-ER) and Weighted Median Estimator (WME) for the sensitivity analysis. These methods calculate the causal estimates using summary statistics obtained from the regression estimates β_{XZj} of the effect of each IVs on the plasma protein level and from the regression estimates β_{YZj} of the effect of each IVs on MS.

4.3.1 Summary statistics methods – an overview

More specifically, IVW, requires that all genetic variants respect the IV assumptions (discussed in previous section), thus assuming that all SNPs are valid instruments[79]. The resulting causal estimate can be interpreted as a weighted regression from the effect estimates of the exposure SNPs on the estimates of the outcome of the same SNPs (without an intercept term)[84].

MR-ER and WME methods allow to detect the causal effect of interest through the simultaneous use of multiple IVs, without requiring that all the instruments satisfy the conditions. Whereas both MR-ER and WME require the untestable assumption c), MR-ER allows assumption b) to be completely relaxed, and WME allows assumption b) to be violated by up to 50% of the instruments.

MR-ER allows more flexibility in terms of weaker requirements than exclusion-restriction criterion (i.e. assumption b): the instrument Z affects Y only through X) and is more robust to potential horizontal pleiotropy (a genetic variant that affects the outcome via a different biological pathway from the exposure under investigation); even if this result in a decrement in the power to detect causal effects. MR-ER method, as IVW, requires the IVs to satisfy the so-called InSIDE assumption, that is the direct pleiotropic effects of the genetic variants on Y have to be distributed independently of the genetic associations with X, in this case it provides a valid test of directional (unbalanced) pleiotropy, and a valid test of the causal null hypothesis. MR-ER uses the InSIDE to estimate the mean pleiotropic

effect and a causal effect adjusted for pleiotropy; the slope estimate is a consistent estimate of the true causal effect while the intercept is the average pleiotropic effect[85].

WME approach allows the IV assumptions to be violated in a more general way for the invalid IVs in respect of MR-ER; in fact, median-based methods are agnostic to the mechanism by which the invalid IVs violate the assumptions. Consistent estimates would be guaranteed if some genetic variants were invalid IVs due to other mechanisms rather than pleiotropy (e.g. LD, population stratification, etc.)[86]. The methods discussed above are synthesized in Table 1.

Method	**Assumptions:**	**Allows:**
IVW	All SNPs must be Valid Instruments InSIDE (Instrument Strength Independent of Direct Effect) Pleiotropy with zero mean across instruments	/
WME	At least 50% of SNPs must be Valid Instruments	Population stratification, Pleiotropy
MR-ER	InSIDE	Directional Pleiotropy

TABLE 1: Assumptions to be respected and bias addressable by each method.

All the analysis described above were performed in R using MendelianRandomization package[87].

4.3.2 Inverse-variance weighted

The causal effect of the exposure on the outcome can be estimated using the *j*th variant as the ratio of the gene-outcome association ($\hat{\Gamma}_j$) and the gene-exposure association estimates[88] ($\hat{\gamma}_j$):

$$\hat{\beta} = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$$

If the IV assumptions are satisfied for genetic variant *j*, then $\Gamma_j = \beta_{\gamma j}$ and the ratio estimate is consistent asymptotically. Furthermore, if the genetic variants are uncorrelated (not in linkage disequilibrium) then the ratio estimates from each genetic variant can be combined into an overall estimate using a formula from the meta-analysis literature:

$$\hat{\beta}_{IVW} = \frac{\sum_j \hat{\gamma}_j \ \sigma_{Yj}^{-2} \hat{\beta}_{Yj}}{\sum_j \hat{\gamma}_j^2 \sigma_{Yj}^{-2}}$$

where σ_{Yj} is the standard error of the gene-outcome association estimate for variant *j*. This is referred to as the inverse-variance weighted (IVW) estimator[79]. Provided that the genetic variants are uncorrelated, the IVW estimate is asymptotically equal to the two-stage least squares estimate commonly used with individual-level data. If all genetic variants satisfy the IV assumptions, then the IVW estimate is a consistent estimate of the causal effect (i.e., it converges to the true value as the sample size increases), as it is a weighted mean of the individual ratio estimates.

4.3.3 Weighted median estimator

The median-based methods have greater robustness to individual genetic variants with strongly outlying causal estimates compared with the inverse-variance weighted and MR-Egger methods. Formally, the simple median method gives a

consistent estimate of the causal effect when at least 50% of the genetic variants are valid instrumental variables (for the weighted median method, when 50% of the weight comes from valid instrumental variables). For the Simple Median Estimator, the estimate is obtained by calculating the ratio causal estimates from each genetic variant $\theta = \beta_Y/\beta_X$, and finding the median estimate.

With the Weighted Median Estimator, the estimate is obtained by calculating the ratio causal estimates and ordering the genetic variants according to the magnitude of their estimates, i.e. $\theta_1 < \theta_2 < \ldots < \theta_j$.

Then calculating the normalized inverse-variance weights for each genetic variant $w_1, w_2, \ldots, w_J$, as:

$$w_j = \frac{\beta_{Xj}^2}{se(\beta_{Yj})^2} \Big/ \sum_{i=1}^{J} \frac{\beta_{Xi}^2}{se(\beta_{Yi})^2}$$

Then finding a *k* such that

$$s_k = \sum_{i=1}^{k} w_i < 0.5 \quad \text{and} \quad s_{k+1} = \sum_{i=1}^{k+1} w_i > 0.5$$

Finally, the weighted median estimate can be calculated by extrapolation as:

$$\theta_{\mathrm{WME}} = \theta_k + (\theta_{k+1} - \theta_k) \times \frac{0.5 - s_k}{s_{k+1} - s_k}$$

The simple median estimate is the same as the weighted median estimate when all the weights are equal. Standard errors for both the simple and weighted median methods are calculated through bootstrapping[86].

4.3.4 MR-Egger regression

"MR-Egger regression" approach to Mendelian Randomization was derived from a method in the meta-analysis literature for the assessment of small-study bias (often called "publication bias")[89]. This performs a weighted linear regression of the gene-outcome coefficients $\hat{\Gamma}_j$ on the gene-exposure coefficients $\hat{\gamma}_j$:

$$\hat{\Gamma}_j = \beta_{0E} + \beta_E \hat{\gamma}_j$$

in which all the $\hat{\gamma}_j$ associations are orientated to be positive (the orientation of the $\hat{\Gamma}_j$ associations should be altered if necessary to match the orientation of the $\hat{\gamma}_j$ parameters), and the weights in the regression are the inverse variances of the gene-outcome associations (σ_{Yj}^{-2}). Reorientation of the variants is performed as the orientation of genetic variants is arbitrary (i.e., estimates can be presented with reference to either the major or minor allele), and different orientations of genetic variants change the estimate of the intercept, as well as the sign and magnitude of the pleiotropic effect of the genetic variant. If there is no intercept term in the regression model, then the MR-Egger slope estimate $\hat{\beta}_E$ will equal the IVW estimate[65].

The value of the intercept term $\hat{\beta}_{0E}$ can be interpreted as an estimate of the average pleiotropic effect across the genetic variants[85]. The pleiotropic effect is the effect of the genetic variant on the outcome that is not mediated via the exposure. An intercept term that differs from zero is indicative of overall directional pleiotropy; that is, pleiotropic effects do not cancel out and the IVW estimate is biased.

MR-Egger regression additionally provides an estimate for the true causal effect $\hat{\beta}_E$ that is consistent even if all genetic variants are invalid due to violation of b), but under a weaker assumption known as the InSIDE (instrument strength independent of direct effect) assumption. If the association of the *j*th genetic variant with the outcome $\Gamma_j = \beta_{Yj} + \alpha_j$ where α_j is the pleiotropic (direct) effect of the variant, then the InSIDE assumption states that the pleiotropic effects α_j must be distributed independently of the instrument strength parameters γ_j[90].

4.3.5 Prioritizing results

One problem arises when applying all the above-mentioned MR methods on a large set of exposures in such an automated way: how to in some way summarize the (potentially different) results coming from the 3 different methods on so many proteins? In fact, while in single specific cases a researcher could explore in detail the possible causes and mechanisms that could lead to different estimates coming from different methods, it's quite clear that this is not possible in a much wider analysis like the one we have conducted here.

In other words, how to be able to prioritize our results basing on the most interesting and reliable ones?

Trying to answer this issue, a couple of possibilities came to our minds. First, the more restrictive one: all 3 methods must be significative for the exposure to be considered causally associated with the disease. Even if this surely leads to more reliable and "secure" results, it is probably a bit too restrictive, eliminating potential interesting findings that maybe did not end up having all significant estimates due to outliers or heterogeneity of IVs.

Another way could be to prioritize the main method (IVW) above the others. This could make sense, but still it would be too unaware of the results coming from the sensitivity analysis, making them probably unnecessary.

The third method we thought of was based on a majority principle: 2 out of 3 significative results would be fine to consider an exposure a significative causal result. Though this seemed to be the best choice so far, we ended up choosing a slightly modified version of it, which in some way consisted in a more technical solution, even if simple: taking the median p-value among the 3 methods. This solution offers a more "statistical" way to discriminate among the results, and also allows, as opposed to the simple majority based one, to end up with a single

estimate and a single p-value to correct for multiple testing still guaranteeing for at least 2 out of 3 of the methods to be significant.

Then, in the end, we ordered results on the basis of the median p-value among the ones resulting from the 3 different methods, and then corrected them for multiple testing applying both Bonferroni and Benjamini-Hochberg (BH) corrections. In this way we ended up with a list of our final prioritized proteins.

4.3.6 Correcting for multiple comparisons

When performing a large number of statistical tests, there's no way to avoid that some will end up having p-values less than 0.05 (the usual significance level) purely by chance, even if all the null hypotheses are really true. The control of the increased type I error (rejecting a true null hypothesis, a "false positive" discovery), when testing simultaneously a family of hypotheses is a central issue in the area of multiple comparisons.

If we, for example, consider a case where we have 20 hypotheses to test at a significance level of 0.05, the probability of observing at least one significant result just due to chance is

$$P(\text{at least one significant result}) = 1 - P(\text{no significant results}) = 1 - (1 - 0.05)^{20} \approx 0.64$$

So, we have a 64% chance of observing at least one significant result, even if all of the tests are actually not significant.

Especially in biological statistical applications, from genomics to many other biology-related fields, it's not unusual for the number of simultaneous tests to be way larger than 20, so that the probability of getting a significant result simply due to chance keeps raising.

This problem has received increasing attention in the last few years and there is no universally accepted approach for dealing with the problem of multiple

comparisons; it is an area of active research, both in the mathematical details and broader epistemological questions.

As for us, in this study we choose to apply two correction methods: the first one, Bonferroni method, is way more restrictive and aims at minimizing the numbers of false positives while the second one, Benjamini-Hochberg correction, is more liberal and powerful and aims at controlling the false discovery rates.

However, both of the methods rely on some assumptions like that the individual tests are independent of each other, which is hardly true in our case. Moreover, our dataset was not generated in an agnostic way, but consists in a set of candidate proteins, selected a priori for their potential involvement in the disease development. For this reason, while in any case we choose to be cautious and more restrictive relying on those corrections methodologies in order to be able to draw more accurate and highly reliable conclusions, it is sure that ours has been a very prudent way, and that the choice of a more liberal approach could be debatable.

Bonferroni

One classical approach to deal with the multiple comparison problem is to control the familywise error rate. Instead of setting the critical P level for significance, or alpha (α), to 0.05, a lower critical value is used. One very common and easy method to do this is the Bonferroni correction, named after the Italian mathematician Carlo Emilio Bonferroni, which sets the significance cut-off at α/n, where n is the number of tests. For example, in the example above, with 20 tests and $\alpha = 0.05$, a null hypothesis is only rejected if the p-value is less than 0.0025.

The Bonferroni correction is appropriate when a single false positive in a set of tests would be problematic. It is mainly useful in situations with a small number of multiple comparisons and few potential significant results, while with a large number of multiple comparisons and many that might be significant, the Bonferroni correction may lead to a high rate of false negatives.

In addition, it assumes that all tests are independent of each other. In practical applications, though, that is often not the case. Depending on the correlation structure of the tests, the Bonferroni correction could be then extremely conservative, leading to raise even more the rate of false negatives.

Benjamini-Hochberg

Therefore, in large-scale multiple testing, an alternative and preferable approach to familywise error rate controlling methods, is controlling instead the false discovery rate (FDR). This is defined as the proportion of false positives among all significant results. The FDR works by estimating some rejection region so that, on average, FDR < α.

One method for controlling the false discovery rate was developed in detail by Yoav Benjamini and Yosef Hochberg[91] and was then named after the two authors.

Briefly, it consists in ordering the individual P values from smallest to largest. The smallest p-value is given a rank of i=1, then next smallest i=2, etc. Then it compares each individual p-value to its BH critical value, $(i/n)Q$, where i is the rank, n is the total number of tests, and Q is the false discovery rate you choose. The largest p-value that has $P<(i/n)Q$ is significant, and all of the p-values smaller than it are also significant. In the end, usually, a corrected p-value is eventually generated. If the BH adjusted P value is smaller than the target false discovery rate, the test is significant. In our case, we considered as reasonable FDR target, due to the exploratory nature of the study, a FDR of 0.10.

4.3.7 Investigating directionality

While we are interested in a putative causal effect of the level of concentration of a protein on the occurrence of MS, we, on the other hand, cannot completely exclude the coexistence of a reverse causal effect, exerted by the disease on the

protein concentration level. This problem of reverse causation is represented by the red arrow from Y to X in figure 7.

The two directions of causality are not necessarily mutually exclusive, and both pathways could play a role in a potential positive feedback loop that would need detailed and direct study.

This situation may create difficulties in the estimation of the magnitude of the causal effects and elucidating the casual direction of the relationship using conventional epidemiological tools is not possible.

To face this condition, in literature some have suggested the use of a bi-directional approach within the framework of Mendelian Randomization[92-94], a way that we chose to follow in this study.

In order to analyze the potential reverse causation, we then decided to perform a bi-directional MR analysis for the proteins showing a significant result in the main causal analysis. On these proteins, we re-run the MR analysis in the opposite direction, thus considering the MS as exposure and protein level as outcome, assessing the possible causal effect of MS on plasma protein level and to this aim, we used the appropriate sets of IVs (i.e. the variants associated with the disease).

Then a significant result from the first analysis will provide evidence of a causal effect of X on Y, and a significant result from the reverse analysis (in addition to the previous, with a disjoint set of independent instruments) will suggest that the causal relationship between X and Y cannot be totally explained in terms of X causing Y, pointing out instead a more complex, and hardly possible to

disentangle, relationship in which both the exposure and the outcome exert some effects on each other.

5. Results

Ut sementem feceris, ita metes.

- Cicero

5.1 Selection of IVs

To be able to select instruments to use in the main MR analysis, we firstly fitted linear mixed-effects kinship models between each protein profile and each SNPs from the set of uncorrelated independent Immunochip genetic variants, using the "lmekin" function implemented in the "coxme" R package, which allows to include in the fitted models random family effects with the input of a kinship correlation matrix. We then estimated association between the same SNPs and a dichotomous phenotype such as MS, using GWAF R package, fitting logistic regression via Generalized Estimation Equation (GEE) using an independence working correlation matrix, with each family acting as a cluster in the robust variance estimation of the genotype effects. Even if considering each family as an independent cluster may not exactly reflect our dataset structure, we considered this approach the more reasonable trade-off between simplification and algorithm efficiency.

A sample of the results obtained by these analyses is reported as example in Table 2, showing association estimates, along with standard errors and p-values, both with the disease and one of the studied proteins of a subset of the analysed SNPs. The table shows estimated associations with a2m_hpa002265 protein and MS, for a sample of 25 SNPs. A similar table with the association estimates for all the 19121 SNPs analysed has been generated for all the 377 proteins.

SNP	Beta MS	SE MS	P-value MS	Beta protein	SE protein	P-value protein
rs1268538	0.29	0.39	0.47	0.03	0.21	0.87
rs12691712	-0.40	0.37	0.28	-0.06	0.18	0.75
rs12692166	-0.34	0.19	0.07	0.11	0.12	0.33
rs12692220	-0.02	0.15	0.88	-0.07	0.13	0.56
rs12692850	0.33	0.22	0.13	0.01	0.16	0.97
rs12693008	0.09	0.18	0.64	-0.04	0.12	0.75
rs12694867	0.14	0.13	0.29	-0.12	0.13	0.35
rs12694912	-0.16	0.16	0.34	0.23	0.11	0.04
rs12695007	0.01	0.16	0.95	-0.19	0.12	0.12
rs12696030	0.10	0.22	0.65	-0.25	0.14	0.06
rs12698020	0.37	0.43	0.39	-0.37	0.23	0.11
rs1269854	-0.05	0.59	0.94	-0.42	0.56	0.45
rs12701626	0.17	0.42	0.68	-0.25	0.27	0.35
rs12703354	-0.66	0.43	0.12	-0.01	0.22	0.96
rs12704637	0.02	0.19	0.92	-0.07	0.13	0.59
rs12705390	-0.21	0.34	0.54	-0.03	0.19	0.89
rs12706382	-0.11	0.39	0.78	0.01	0.24	0.97
rs12706940	-0.03	0.14	0.85	0.05	0.11	0.69
rs12709148	-0.12	0.15	0.43	0.17	0.12	0.16
rs12710675	0.12	0.11	0.28	0.07	0.12	0.56
rs12711517	-0.24	0.17	0.16	-0.15	0.12	0.19
rs12712078	0.08	0.16	0.63	0.00	0.12	0.97
rs12712691	-0.02	0.13	0.87	-0.31	0.11	0.01
rs12712696	-0.18	0.32	0.58	-0.05	0.16	0.78
rs12712880	0.04	0.21	0.84	-0.41	0.13	0.002

TABLE 2: Betas, standard errors and p-values for associations with MS and with a2m_hpa002265 protein. Here shown a sample of 25 SNPs (out of the 19121 analysed).

For the main MR analysis, we needed to select for each protein only the variants significantly associated with it. Therefore, we chose an a priori significance threshold of $p<5x10^{-4}$ to identify and select significant associated SNPs with each protein.

In Table 15 in the appendix, is shown the final number of SNPs being selected as Instrumental Variables in this way for each protein. The protein ending up with the higher number of IVs was cp_hpa001834, having 43 genetic variants associated with a p-value$<5x10^{-4}$. Proteins ending up with less than 3 variants couldn't be analysed using the previously mentioned MR summary statistics methods and were therefore excluded by the final analysis. Only three proteins among the whole set, ending up with a non-sufficient number of IVs, were excluded for this reason from the consequent MR stage: c1orf106kif21b_hpa027499, chgb_hpa012602 and csda_hpa034838 (see Table 3). For the remaining proteins, the mean number of IVs is 16.73, median is 17.

Antibody ID	N° of IVs
c1orf106kif21b_hpa027499	2
chgb_hpa012602	2
csda_hpa034838	1

TABLE 3: Proteins ending up with less than 3 variants associated with a p-val$<5x10^{-4}$ and therefore excluded by the consequent MR analysis.

5.2 Mendelian Randomization analysis

In the main MR causal analyses, we applied three different MR methods, Inverse-Variance Weighted (IVW) as primary method for the main causal analysis and MR Egger regression (MR-ER) and Weighted Median Estimator (WME) for the sensitivity analysis, on each of our proteins. In table 4-6 we show an example of the results obtained for a2m_hpa002265, ablm2_hpa035808 and ace2_hpa000288 proteins. Causal estimates are provided as logOdds Ratio per 1 standard deviation increase along with standard errors, confidence intervals (95%) and uncorrected p-values. For a2m_hpa002265, only Egger Regression method showed significant estimates at the nominal level. Ablm2_hpa035808 and ace2_hpa000288 showed no significant results for any of the applied methods.

Along the tables with results we were able to generate graphs of the obtained estimates thanks to the "mr_plot" function implemented in the "MendelianRandomization" package. The function generates a scatter plot of the genetic associations with the protein and with the disease and compares the causal estimates obtained via the different MR methods applied. Being MR-ER fitted to the summary data when the SNP-exposure associations are oriented in the positive direction, in order to make it easier to interpret, the SNPs have been oriented before generating the plots.

The graphs for the same 3 proteins are shown here as examples (Figures 8-10) for sake of space and text readability, while it's important to note that tables and graphs like those presented were generated for all the 374 analysed proteins.

METHOD	Beta	SE	Lower CI	Upper CI	P-value
IVW	-0.09	0.11	-0.30	0.11	0.37
WME	-0.17	0.15	-0.47	0.13	0.26
MR-EGGER	-0.44	0.24	-0.91	0.02	0.06
(intercept)	0.23	0.14	-0.05	0.52	0.11

TABLE 4: Betas, standard errors, 95% confidence intervals and uncorrected p-values resulting from Inverse-variance Weighted, Weighted Median Estimator and Egger Regression methods for a2m_hpa002265 protein.

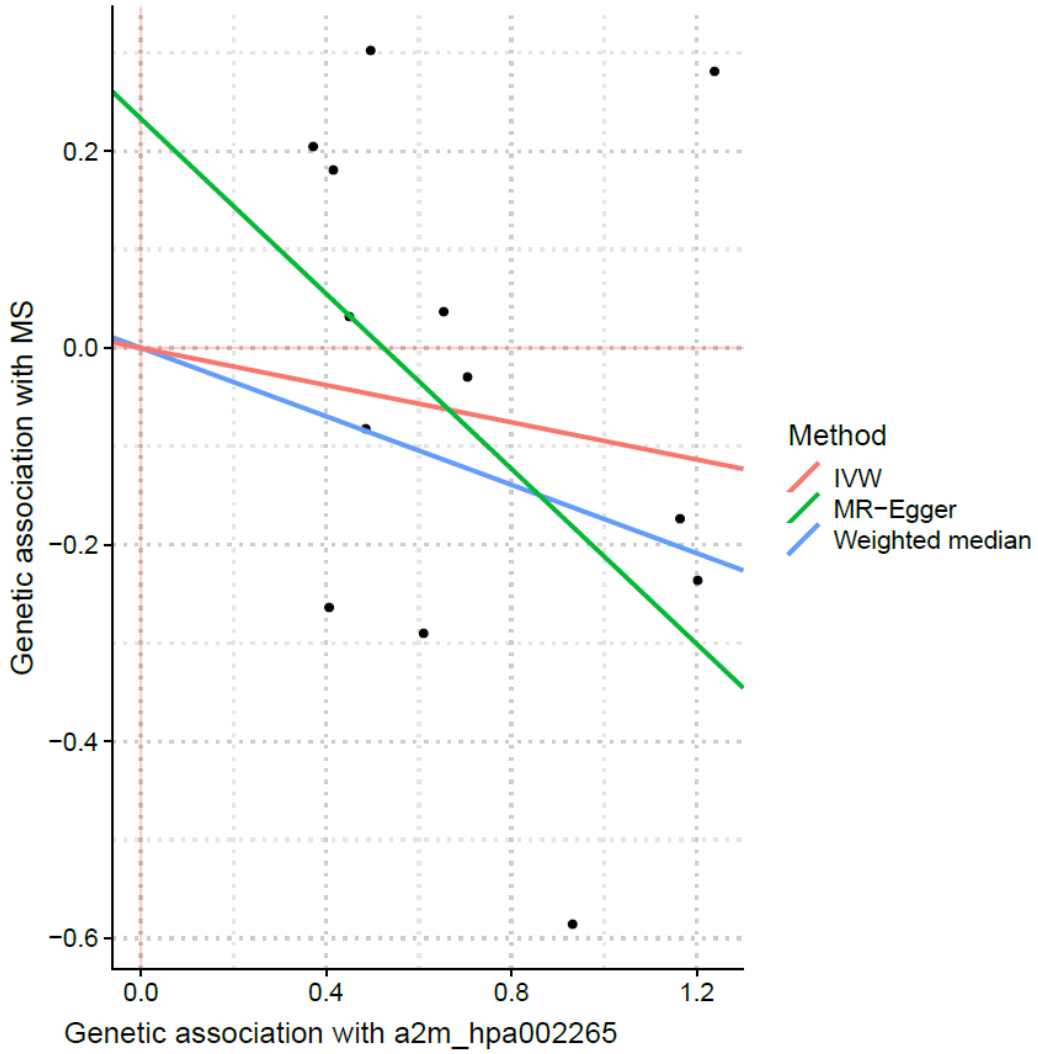


FIGURE 8: Plot of the betas of association with MS (y axis) and with the protein (x axis, positively oriented). Different coloured lines represent causal estimates obtained by the three different applied methods. Each dot is a single genetic variant.

METHOD	Beta	SE	Lower CI	Upper CI	P-value
IVW	-0.18	0.21	-0.59	0.23	0.39
WME	-0.22	0.24	-0.70	0.25	0.36
MR-EGGER	0.27	1.14	-1.96	2.50	0.81
(intercept)	-0.23	0.57	-1.36	0.90	0.69

TABLE 5: Betas, standard errors, 95% confidence intervals and uncorrected p-values resulting from Inverse-variance Weighted, Weighted Median Estimator and Egger Regression methods for, ablm2_hpa035808 protein.

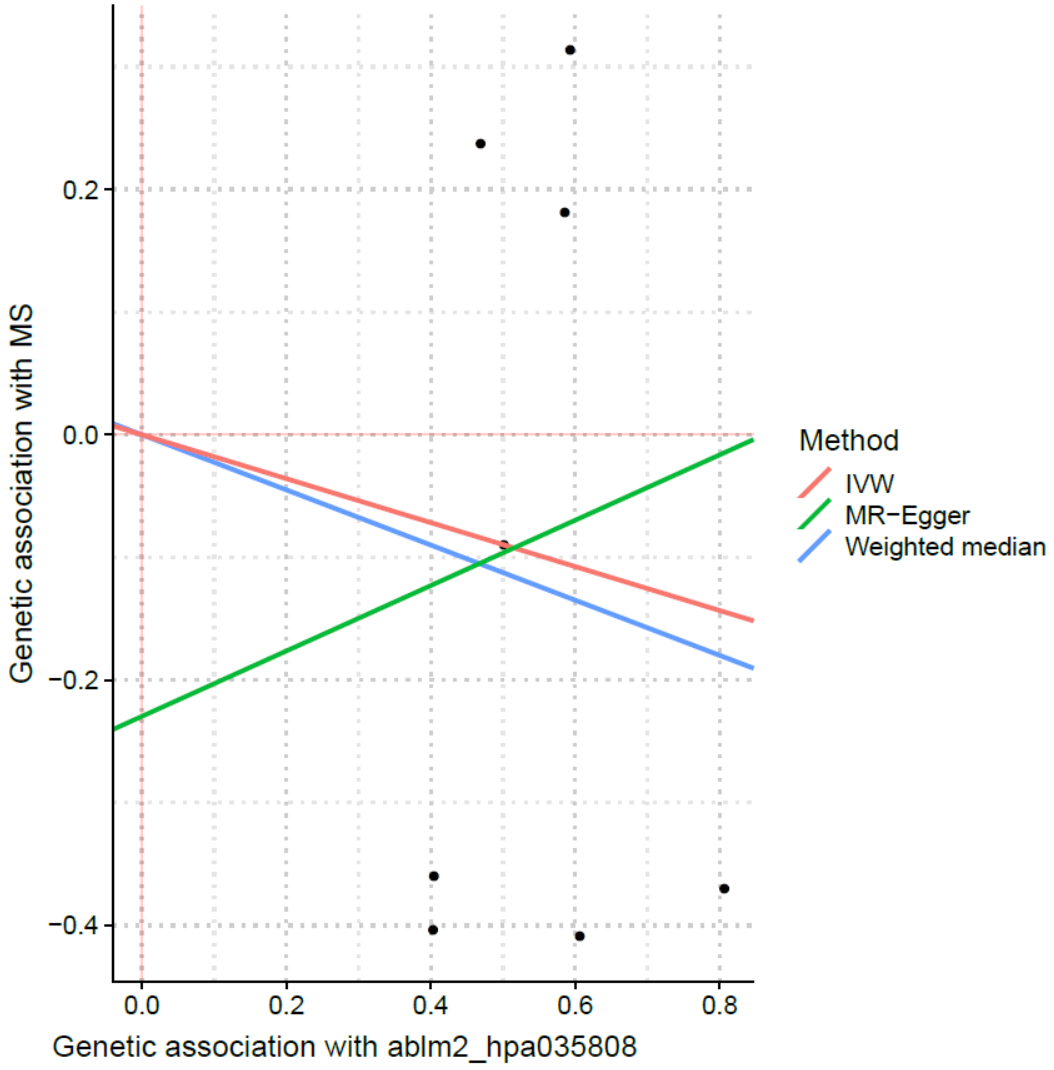


FIGURE 9: Plot of the betas of association with MS (y axis) and with the protein (x axis, positively oriented). Different coloured lines represent causal estimates obtained by the three different applied methods. Each dot is a single genetic variant.

METHOD	Beta	SE	Lower CI	Upper CI	P-value
IVW	0.09	0.10	-0.10	0.28	0.37
WME	0.13	0.12	-0.11	0.38	0.29
MR-EGGER	0.35	0.32	-0.28	0.97	0.27
(intercept)	-0.14	0.16	-0.44	0.17	0.39

TABLE 6: Betas, standard errors, 95% confidence intervals and uncorrected p-values resulting from Inverse-variance Weighted, Weighted Median Estimator and Egger Regression methods for ace2_hpa000288 protein.

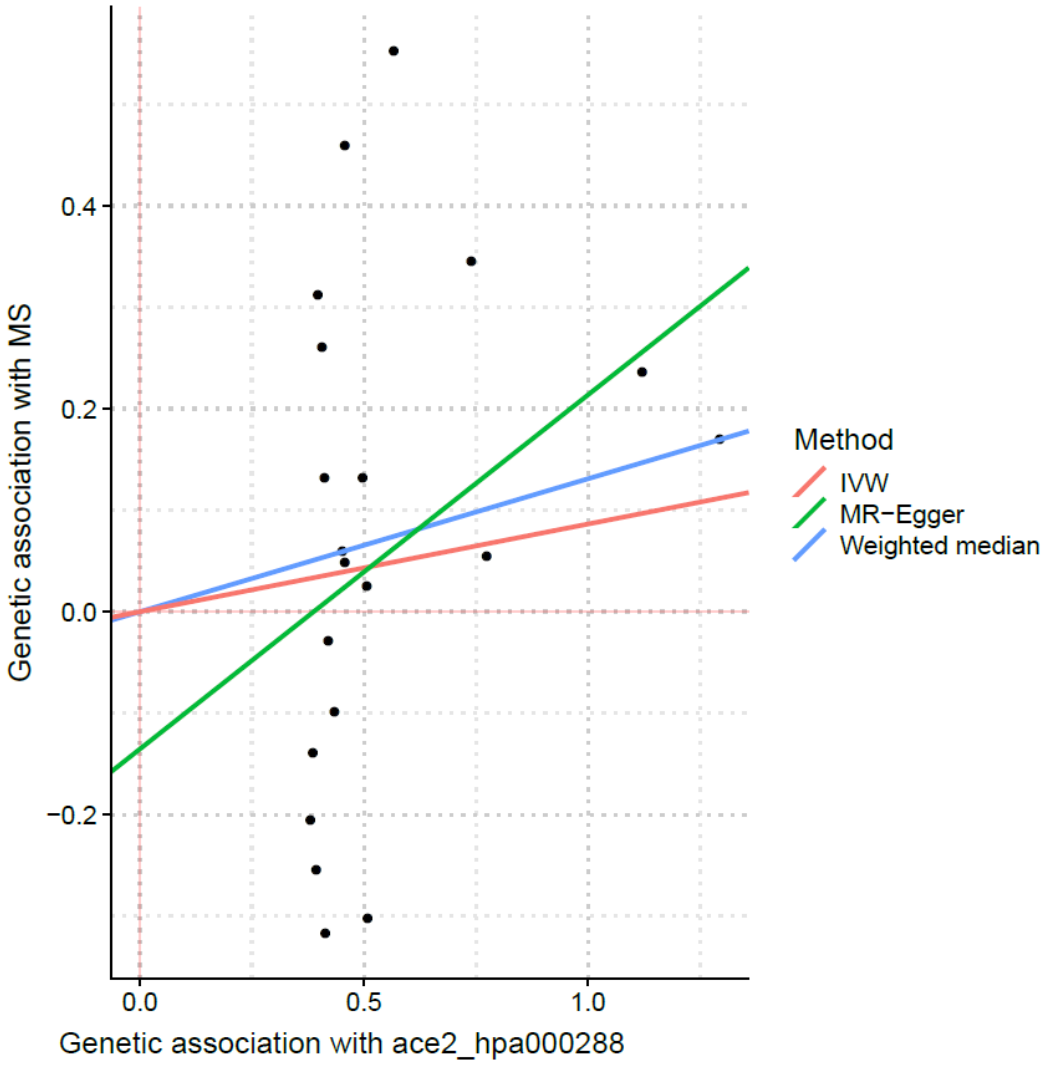


FIGURE 10: Plot of the betas of association with MS (y axis) and with the protein (x axis, positively oriented). Different coloured lines represent causal estimates obtained by the three different applied methods. Each dot is a single genetic variant.

5.3 Prioritizing results

In order to be able to prioritize our findings among the very long list of results, we decided to summarise the estimates obtained for each protein taking the median p-value among the 3 methods. We ended up with a single estimate and a single p-value for each protein, still guaranteeing for at least 2 out of 3 of the methods being significant, which we were able to order and correct for multiple testing using both a more restrictive Bonferroni and a more permissive Benjamini-Hochberg correction.

In this way we ended up with a list of our final prioritized proteins. In Table 7 we show a sample of the results, in particular the 40 proteins showing a significant p-value before the correction. After correction for multiple testing, only 3 proteins showed significant results both with Bonferroni (adj. p-value<0.05) and BH (adj. p-value<0.10) corrections: mobp_hpa035152 (Bonf.: 0.01 ; BH: 0.01), zmynd19_hpa020642 (Bonf.: 0.04 ; BH: 0.01) and kiaa0494_hpa011224 (Bonf.: 0.04 ; BH: 0.01) (See Table 8).

PROTEIN	METHOD	Beta	SE	Lower CI	Upper CI	P-value	Bonf.	BH
mobp_hpa035152	IVW	0.24	0.06	0.12	0.35	$3.08e^{-5}$	0.01	0.01
zmynd19_hpa020642	WME	0.37	0.10	0.18	0.56	0.0001	0.04	0.01
kiaa0494_hpa011224	MR-Egger	-0.69	0.18	-1.05	-0.34	0.0001	0.04	0.01
bcl6_hpa004899	WME	-0.28	0.09	-0.45	-0.10	0.002	0.68	0.17
enw1_hpa003407	WME	-0.28	0.09	-0.46	-0.10	0.003	1	0.21
gnl2_hpa027163	WME	0.45	0.16	0.14	0.75	0.004	1	0.24

PROTEIN	METHOD	Beta	SE	Lower CI	Upper CI	P-value	Bonf.	BH
nalcn_hpa031889	WME	0.37	0.13	0.11	0.62	0.005	1	0.25
dusp8_hpa020071	IVW	-0.37	0.13	-0.63	-0.11	0.01	1	0.28
eif3h_hpa023117	WME	-0.52	0.19	-0.89	-0.14	0.01	1	0.30
eomes_hpa028896	WME	0.42	0.16	0.11	0.74	0.01	1	0.30
plek_hpa031838	IVW	-0.17	0.06	-0.30	-0.04	0.01	1	0.30
sh3bgrl3_hpa030848	IVW	0.18	0.07	0.04	0.31	0.01	1	0.30
ndfip1_hpa009682	IVW	0.27	0.11	0.06	0.48	0.01	1	0.30
tf_hpa001527	WME	-0.29	0.12	-0.52	-0.06	0.01	1	0.30
fbln1_hpa001642	WME	0.38	0.15	0.08	0.69	0.01	1	0.30
tnfsf13_hpa004863	WME	-0.27	0.11	-0.49	-0.06	0.01	1	0.30
gfi1b_hpa007012	WME	-0.44	0.18	-0.79	-0.09	0.01	1	0.30
arhgef3_hpa034715	WME	0.20	0.08	0.04	0.36	0.02	1	0.34
il4_hpa007714	MR-Egger	0.73	0.31	0.13	1.32	0.02	1	0.34
nos2a_hpa003871	IVW	-0.23	0.10	-0.42	-0.04	0.02	1	0.34
hyls1_hpa041210	MR-Egger	-0.57	0.24	-1.04	-0.09	0.02	1	0.34
c1qa_hpa002350	WME	0.21	0.09	0.03	0.39	0.02	1	0.36
stx11_hpa007992	WME	-0.28	0.12	-0.53	-0.04	0.02	1	0.36
angi_hpa036018	MR-Egger	0.63	0.28	0.08	1.17	0.02	1	0.36

PROTEIN	METHOD	Beta	SE	Lower CI	Upper CI	P-value	Bonf.	BH
smyd2_hpa029023	MR-Egger	0.73	0.32	0.09	1.37	0.02	1	0.36
gda_hpa024099	WME	0.47	0.21	0.05	0.89	0.03	1	0.36
il16_hpa018467	WME	0.30	0.14	0.03	0.57	0.03	1	0.36
xpc_hpa035706	WME	-0.48	0.22	-0.92	-0.05	0.03	1	0.36
trm13_hpa028494	MR-Egger	0.51	0.23	0.05	0.97	0.03	1	0.36
pdgfb_hpa011972	IVW	0.24	0.11	0.02	0.45	0.03	1	0.36
heatr3_hpa041990	IVW	-0.18	0.08	-0.34	-0.02	0.03	1	0.36
prickle4_hpa031240	IVW	-0.23	0.11	-0.44	-0.02	0.03	1	0.36
taf8_hpa031730	MR-Egger	0.24	0.11	0.02	0.46	0.03	1	0.36
dsg1_hpa022128	IVW	0.14	0.07	0.01	0.27	0.03	1	0.37
slc30a7_hpa018034	MR-Egger	0.94	0.45	0.05	1.82	0.04	1	0.40
rpesp_hpa029595	IVW	0.25	0.12	0.01	0.48	0.04	1	0.40
dlst_hpa003010	MR-Egger	-0.69	0.34	-1.36	-0.03	0.04	1	0.42
dars_hpa024079	IVW	-0.20	0.10	-0.39	-0.01	0.04	1	0.42
taf8_hpa031734	IVW	-0.18	0.09	-0.36	0.00	0.04	1	0.42
btn3a1_hpa012565	IVW	0.16	0.08	0.00	0.31	0.045	1	0.42

TABLE 7: MR “median” methods, betas, standard errors, 95% confidence intervals and p-values (uncorrected, Bonferroni correction, Benjamini-Hochberg correction) of the 40 proteins showing significant p-values before corrections.

PROTEIN	METHOD	Beta	SE	Lower CI	Upper CI	P-value	Bonf.	BH
mobp_hpa035152	IVW	0.24	0.06	0.12	0.35	$3.08e^{-5}$	0.01	0.01
zmynd19_hpa020642	WME	0.37	0.10	0.18	0.56	0.0001	0.04	0.01
kiaa0494_hpa011224	MR-Egger	-0.69	0.18	-1.05	-0.34	0.0001	0.04	0.01

TABLE 8: MR "median" methods, betas, standard errors, 95% confidence intervals and p-values (uncorrected, Bonferroni correction, Benjamini-Hochberg correction) of the 3 proteins showing significant p-values after correction for multiple testing.

The protein that obtained the lowest adjusted p-value, under Bonferroni correction, is mobp_hpa035152.

In Table 9 and Figure 11 the causal estimates for this protein obtained by the three MR methods applied are reported with SEs, Cis and unadjusted p-values. All the methods showed significant causal estimates, betas all pointing to a consistent detrimental effect of an increased level of the protein toward the disease. Intercept from MR-Egger method is significantly different from 0 suggesting potential pleiotropy among the instrumental variables.

METHOD	Beta	SE	Lower CI	Upper CI	P-value
WME	0.31	0.09	0.12	0.49	0.001
IVW	0.24	0.06	0.12	0.35	$3.08e^{-5}$
MR-EGGER	0.53	0.12	0.30	0.77	$5.51e^{-6}$
(intercept)	-0.21	0.07	-0.36	-0.07	0.005

TABLE 9: MR methods, betas, standard errors, 95% confidence intervals and uncorrected p-values for mobp_hpa035152.

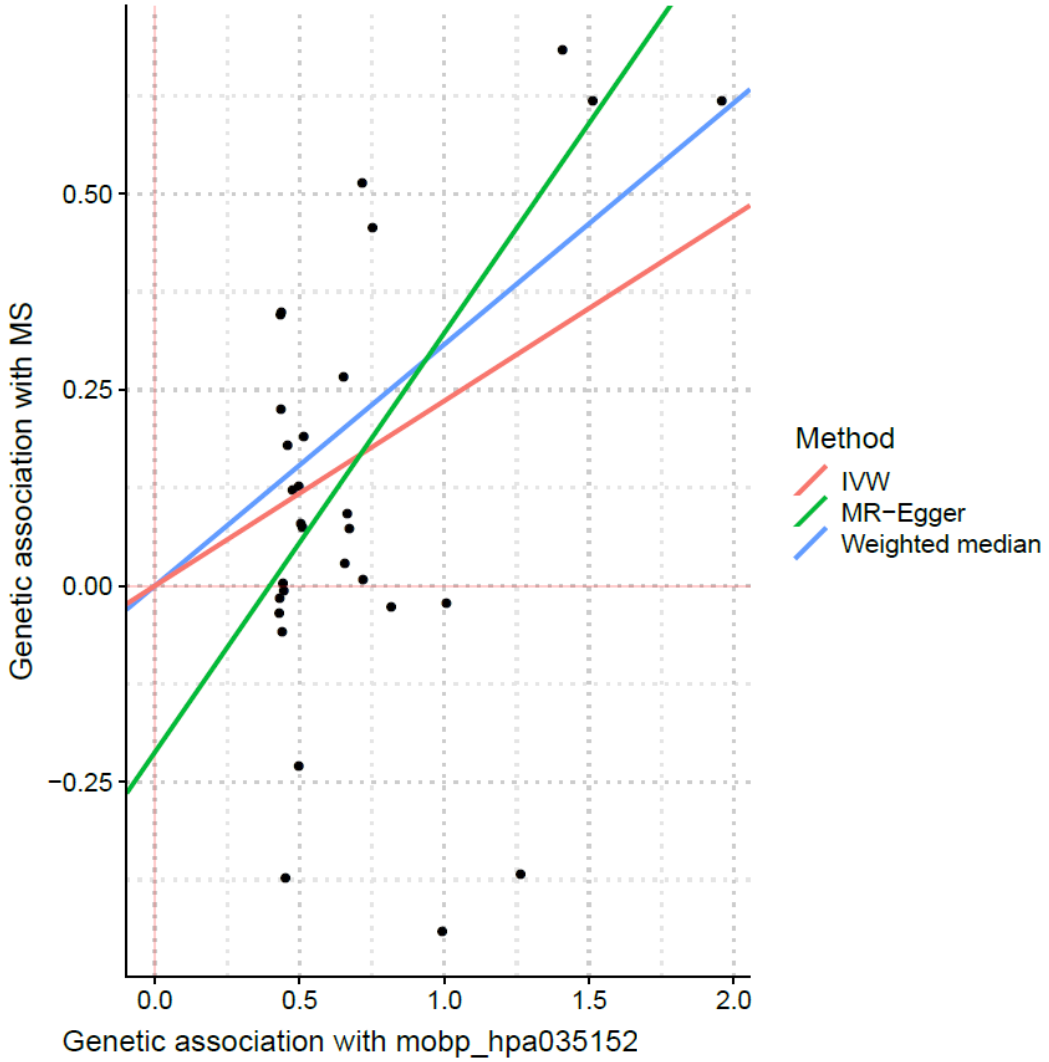


FIGURE 11: Plot of the betas of association with MS (y axis) and with mobp_hpa035152 (x axis, positively oriented). Different coloured lines represent causal estimates obtained by the three different applied methods. Each dot is a single genetic variant.

For the zmynd19_hpa020642 protein (Table 10, Figure 12) IVW and WME methods gave significant results and concordant causal estimates (log Odds for IVW: 0.32, WME: 0.37), IVW being the one with the lowest p-value overall, while MR-Egger method gave highly non-significant result (p-value: 0.97).

METHOD	Beta	SE	Lower CI	Upper CI	P-value
WME	0.37	0.10	0.18	0.56	0.0001
IVW	0.32	0.06	0.21	0.44	$2.31e^{-8}$
MR-EGGER	-0.01	0.21	-0.42	0.40	0.97
(intercept)	0.20	0.12	-0.04	0.43	0.10

TABLE 10: MR methods, betas, standard errors, 95% confidence intervals and uncorrected p-values for zmynd19_hpa020642.

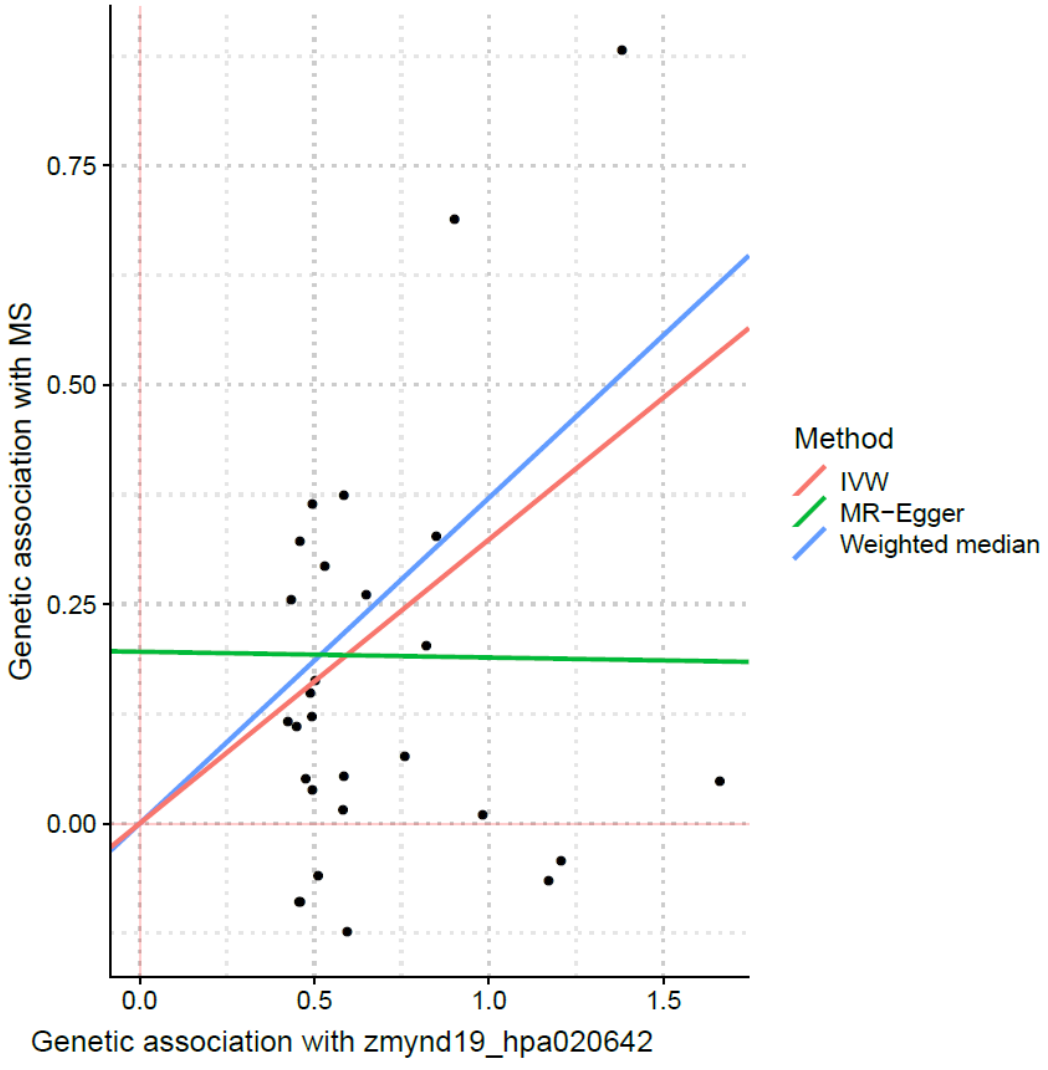


FIGURE 12: Plot of the betas of association with MS (y axis) and with zmynd19_hpa020642 (x axis, positively oriented). Different coloured lines represent causal estimates obtained by the three different applied methods. Each dot is a single genetic variant.

For kiaa0494_hpa011224 all the 3 methods showed significant p-values, IVW showing the lowest one, and concordant causal estimates in a protective direction: it seems that an increased level of the protein may cause a decreased risk of developing the disease (Table 11, Figure 13). The intercept from the MR-Egger method is significantly different from 0, too, suggesting a possible presence of pleiotropy among the variants.

METHOD	Beta	SE	Lower CI	Upper CI	P-value
WME	-0.32	0.10	-0.51	-0.13	0.001
IVW	-0.27	0.07	-0.41	-0.14	$5.18e^{-9}$
MR-EGGER	-0.69	0.18	-1.05	-0.34	0.0001
(intercept)	0.25	0.10	0.05	0.45	0.01

TABLE 11: MR methods, betas, standard errors, 95% confidence intervals and uncorrected p-values for kiaa0494_hpa011224.

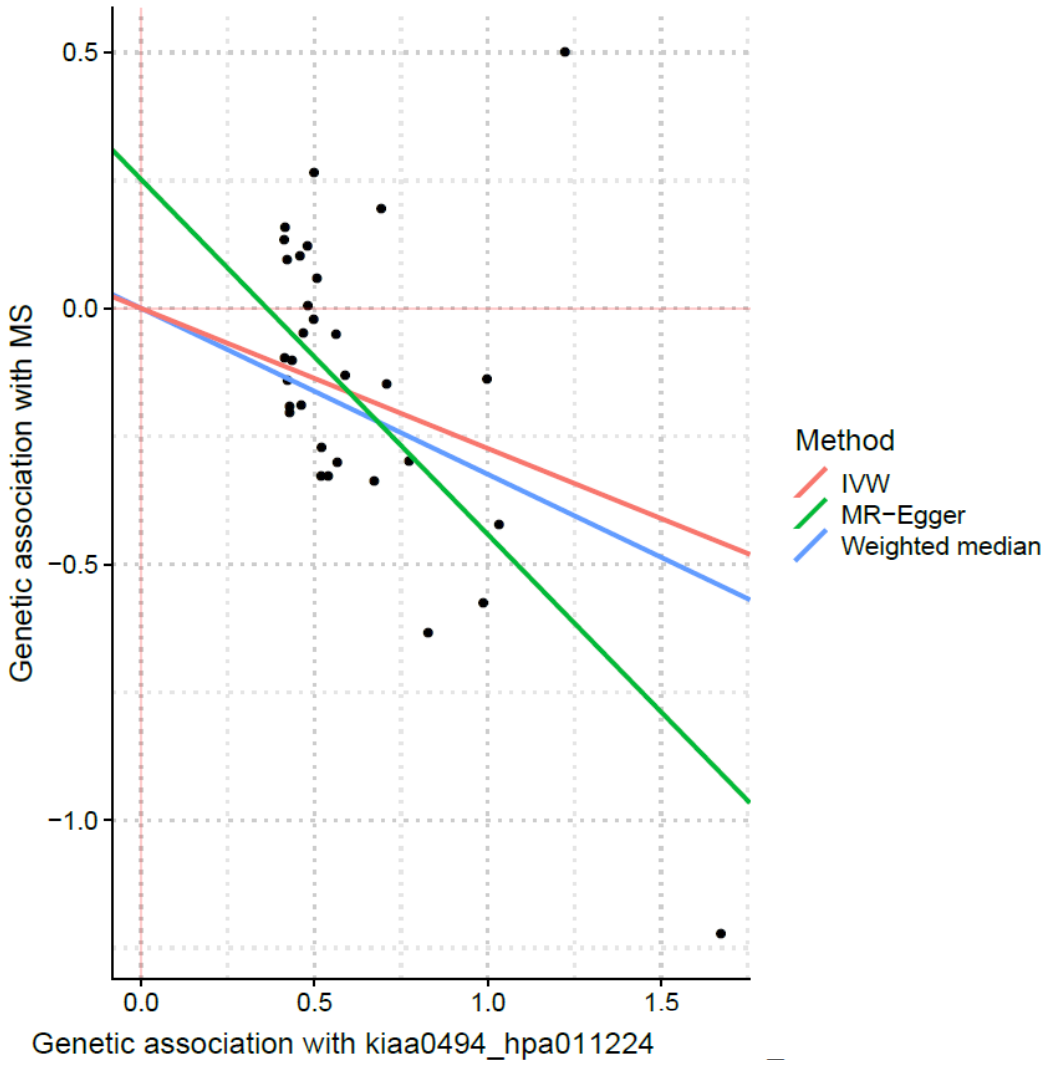


FIGURE 13: Plot of the betas of association with MS (y axis) and with kiaa0494_hpa011224 (x axis, positively oriented). Different coloured lines represent causal estimates obtained by the three different applied methods. Each dot is a single genetic variant.

5.4 Testing directionality

For the 3 proteins that showed significant results in the main MR analysis, we then performed the whole set of analysis as a further test for a possible causal effect in the reverse direction, by using each protein level as outcome and MS as risk factor. Non-significant causal estimates would be regarded as a confirmation of the previous results, and the proteins involved would be considered reliable causal actors toward the disease; significant estimates instead would suggest some potential presence of reverse-causation loops in the pathway and would lead to the exclusion of the proteins involved from the causal candidates.

Results of this analysis are reported in Table 12-14.

METHOD	Beta	SE	Lower CI	Upper CI	P-value
WME	0.04	0.04	-0.03	0.11	0.31
IVW	0.03	0.03	-0.02	0.08	0.30
MR-EGGER	0.06	0.08	-0.09	0.22	0.44
(intercept)	-0.02	0.04	-0.11	0.07	0.65

TABLE 12: MR methods, betas, standard errors, 95% confidence intervals and p-values for mobp_hpa035152 (reverse-causation analysis).

The reverse-direction analysis on mobp_hpa035152 showed no significant estimates for any of the applied methods (IVW p-value: 0.30, WME p-value: 0.31, MR-Egger p-value: 0.44), suggesting no direct effects of the disease on the protein levels.

This empowers our previous results and confirms the protein as a reliable causal candidate for Multiple Sclerosis.

METHOD	Beta	SE	Lower CI	Upper CI	P-value
WME	0.10	0.04	0.03	0.18	0.01
IVW	0.08	0.03	0.02	0.14	0.01
MR-EGGER	-0.02	0.09	-0.19	0.15	0.82
(intercept)	0.06	0.05	-0.04	0.15	0.23

TABLE 13: MR methods, betas, standard errors, 95% confidence intervals and p-values for zmynd19_hpa020642 (reverse-causation analysis).

Testing the causal pathway leading from MS to zmynd19_hpa020642 ended up in a significant causal estimate for 2 out of the 3 methods, IVW and WME, with WME being the "median" method with a p-value of 0.0074 (approximated at 0.01 in Table 13). Only MR-Egger showed a non-significant result. This analysis strongly suggests the presence of some potential reverse effect exerted by the disease itself on the protein levels, confusing and masquerading somehow the real biological pathway and making it impossible to exactly disentangle the direction of the relationship between the two. For this reason, this protein had to be removed from the list of reliable causal candidates of MS.

METHOD	Beta	SE	Lower CI	Upper CI	P-value
WME	-0.03	0.04	-0.10	0.04	0.45
IVW	-0.003	0.02	-0.05	0.05	0.89
MR-EGGER	-0.05	0.07	-0.20	0.09	0.49
(intercept)	0.03	0.04	-0.05	0.11	0.49

TABLE 14: MR methods, betas, standard errors, 95% confidence intervals and p-values for kiaa0494_hpa011224 (reverse-causation analysis).

The reverse-direction analysis on kiaa0494_hpa011224 showed no significant estimates for any of the applied methods (IVW p-value: 0.89, WME p-value: 0.45,

MR-Egger p-value: 0.49), suggesting no direct effects of the disease on the protein levels.

This empowers our previous results and confirms the protein as a reliable causal candidate for Multiple Sclerosis.

6. Conclusions

We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions.

- R. Fisher

6.1 MOBP - Myelin-associated oligodendrocyte basic protein

MOBP, or better Myelin-Associated Oligodendrocyte Basic Protein, is a protein encoded by the MOBP gene, which is a 62,300 bases long protein coding gene situated on the plus strand of chromosome 3 (Fig. 14).

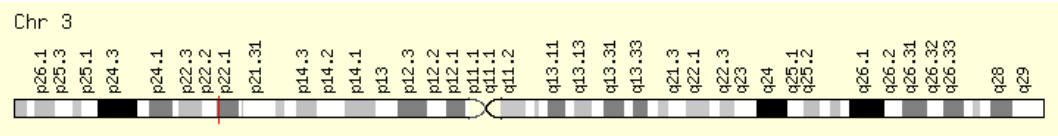


FIGURE 14: MOBP Gene in genomic location: bands according to Ensembl, locations according to GeneLoc.

Diseases associated with MOBP include Substance Abuse and, more interestingly for us, Multiple Sclerosis.

The protein is 183 amino acids long, and its molecular function so far has been individuated, as reported in GeneCards website, in playing “a role in compacting or stabilizing the myelin sheath, possibly by binding the negatively charged acidic phospholipids of the cytoplasmic membrane”. As it appears clearly, it therefore stands as a very plausible candidate for being deeply involved in the aetiology of this disease.

From our analysis, it showed a final Odds Ratio of 1.27 (beta: 0.24) towards the disease from Inverse-Variance Weighted method, confirmed by a highly significant p-value, suggesting how an increase in the level of the protein could end up in an increased risk of developing the disease.

This result could somehow confirm previous literature findings that have shown how MOBP region, and in particular some specific MOBP protein epitopes, is potentially highly relevant for T-cell reactivity against it to MS, both in mice and in humans, being associated with the emergence of many MS/EAE symptoms like intense perivascular and parenchymal infiltrations, widespread demyelination, axonal loss, and remarkable optic neuritis, and can be considered a primary target antigen in MS[95, 96]. From this point of view, higher levels of MOBP could lead to an increased pathogenic autoimmune response by the targeting autoimmune T-cells and induce more severe symptoms emergence.

This seems to fit with the results from a study by Holz *et al*.[97] in which Peripheral Blood Lymphocytes obtained from patients with relapsing/remitting multiple sclerosis mount a proliferative response to human MOBP , showing its association with Multiple Sclerosis.

Our finding seems to confirm and enforce these previous studies, focusing only on human cohorts and adopting a statistically sophisticated method to properly assess a causal link between the protein and the disease, suggesting further specific analysis in this direction in order to better evaluate and refine the understanding of its role in the development of the disease and, hopefully, studying potential useful interventions on it.

6.2 KIAA0494 (EFCAB14) - EF-hand calcium-binding domain-containing protein 14

KIAA0494, from now on EFCAB14, or better EF-Hand Calcium Binding Domain 14, is a protein encoded by EFCAB14, a Protein Coding gene 43,906 bases long situated on the minus strand of chromosome 1. (Fig. 15)

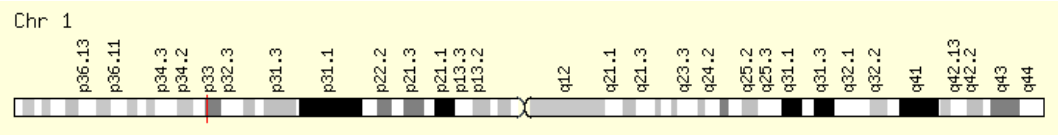


FIGURE 15: KIAA0494 Gene in genomic location: bands according to Ensembl, locations according to GeneLoc.

This gene is mainly associated with calcium ion binding and metal ion binding, and the main phenotypes that have been associated with it are Chronic Hepatitis C infection, obese body mass index status and urate measurement.

The protein is 495 amino acids long, and interestingly has not any specific molecular function or pathway enrichment annotations and could therefore be considered as a novel finding in regard to Multiple Sclerosis. On the other hand, though, understanding more deeply its role in the disease development and exploring how it could be involved in Multiple Sclerosis aetiology is everything but a simple task and would surely require some additional specifically planned studies.

In our study the Odds Ratio from the "median" method for this protein, which was MR-Egger method, resulted to be 0.50 (beta: -0.69), with a p-value of $1e^{-04}$, showing therefore a protective effect toward the disease. It seems in fact that an increased level of EFCAB14 leads to a lower risk of developing Multiple Sclerosis, as confirmed by the concordant results from all the methods applied.

It is nowadays well known that Ca^{2+} overload is one of the main processes that could lead to neurons damage and death, and many studies have shown the

importance of this aspect eventually bringing to the emergence of the so-called "Ca^{2+} *theory of neurodegeneration"* [98]. The key signaling role and involvement of Calcium ion in many different intracellular and extracellular processes, from synaptic activity to cell-cell communication and adhesion, has been widely described and reviewed. Especially in the brain, calcium plays a fundamental role in controlling the synaptic activity and memory formation, and properly controlled homeostasis of calcium signaling not only supports normal brain physiology but also maintains neuronal integrity and long-term cell survival. On the other hand, Calcium deregulation can lead instead to neurodegeneration via complex and diverse mechanisms involved in selective neuronal impairments and death[99]. Since the first evidences, it was proposed that the pharmacological blockade Ca^{2+} overload could rescue neurons from death, and therefore identifying excess Ca^{2+} sources in neurodegenerative diseases has been and still is the object of great interest and attention[100].

Another interesting study as shown that increased levels of retinal calcium and calpain activation are early events in autoimmune optic neuritis[101], which is a common manifestation of Multiple Sclerosis itself.

Deregulation of brain metal ion homeostasis has also emerged as a critical common feature across different neurodegenerative diseases and cumulating evidence points to pathological changes in the neuronal balance of metal ions such as zinc, calcium, iron and copper in these diseases[102, 103].

Trying to speculate a bit over what's hidden behind our results, we could say that our finding is driven by a deep involvement of EFCAB14 in the active regulation (and deregulation) of calcium and other metal ions and possibly in maintaining the homeostasis in brain cells. Its known role in calcium and metal ion binding could be somehow related to the regulation of these potentially noxious elements, binding "wandering" ions in excess and in this way blocking their way towards accumulation and consequent involvement in neurodegenerative processes, explaining the protective effect of increased levels of the protein against Multiple

Sclerosis that we found in our study. Without further specific analysis that would explore molecular mechanisms in which this protein acts, though, these remain pure hypothesis.

6.3 Summing up

In this study we applied a Mendelian Randomization causal inference approach, integrating plasma protein and genotyping data on 20 multiplex Sardinian families, in order to identify potential causal mechanisms between the plasma levels of a large set of 377 candidate proteins and MS, with the aim of exploring MS pathogenesis. Many genome-wide association studies (GWASs) have studied the association between millions of genetic variants and many outcomes, thus identifying many genetic signals statistically associated with many risk factors (e.g. obesity, biomarkers, gene expression level) that nowadays can be used to investigate the causal association between these risk factors and the disease of interest[4, 104]. Causal questions are, in fact, what motivate, more or less explicitly, the vast majority of statistical and epidemiological studies. Unfortunately, causal conclusion might not be warranted by the data; the causal interpretation may be flawed and the risk estimates may be biased, making difficult to distinguish between causal associations from associations that arises from confounding or reverse causation[88].

The approach we applied overcomes the limitations of both residual confounding, which can bias the relationship between risk factors and disease in observational studies, and reverse causation, particularly likely here since the level of plasma protein may be affected by the disease itself. In this context, the use of IVW in association with two sensitivity analyses, i.e. MR-ER and WME, has helped to face these limitations and to assess robustness of causal findings to different sets of assumptions regarding independence of IVs and pleiotropy. We further performed the MR analysis in the opposite direction, in an approach usually referred to as bidirectional MR; this allowed us to test specific reverse causation effects in order to confirm or refute causal findings revealed in the first part of the analysis.

We did this by applying a custom automated-routine, coded in R, that allowed us to run all the MR analyses for the whole large number of proteins we had, without

having to deal with each single analysis in turn, which would have been hugely time-expensive.

Nonetheless, this approach presented us some other problems to tackle: how to manage the vast quantity of results, possibly discordant between the different MR methods applied, and, of course, the need to deal with multiple comparisons bias. To handle the first problem, we then chose to select a "median" method for each protein as described in the method section, and we then ordered and corrected all our results for multiple testing, both with Bonferroni and BH corrections, handling the second one in a quite cautious way; we preferred to have highly reliable results even at the cost of losing some potential interesting information (false negative results).

In the end, 3 proteins showed significant results with both corrections applied, in particular MOBP, ZMYND19 and EFCAB14.

Following the bidirectional analysis though, ZMYND19 showed a significant result in the reverse-direction too, suggesting some reverse causation effect. It seems that, in this case, the disease itself could influence the level of this protein in plasma.

The final and most interesting findings in the end are therefore MOBP and EFCAB14, that we extensively treated in the previous chapters.

Whereas MR methods are typically applied to high-level exposures, such as obesity and blood pressure[105-107], in our study we used standard MR methods to identify concentration of specific plasma proteins causally related to MS. Our proposed application on protein biomarkers instrumented by SNPs has the advantage of being the protein a close consequence of DNA sequence variations. We have investigated both SNPs in the encoding gene or acting *in cis* and SNPs in the distal region thus involved in gene expression regulation *in trans.* Even if these latter have an high probability to assert pleiotropic effect they allow to investigate the functional meaning and the effect of the statistically associated variants located outside the protein-coding region[4].

We based our analysis on a restricted number of subjects belonging to pedigree and hence correlated, and in particular to a limited number of cases, due to the very structure of our starting dataset and availability of a set of plasmatic concentration of certain candidate proteins and Immunochip data as source of genotyping data to identify IVs.

This led to some difficult choices, as the threshold choice for the selection of the instruments to be used in the consequent MR analyses. In the end, being our set of proteins already a selection of candidate ones and coming the genotyping data from an Immunochip array and not from a GWAS study, we chose an a priori significance threshold of $p<5x10^{-4}$ to identify and select significant associated SNPs with each protein considering this one a reasonable trade-off between being more restrictive and having more reliable instruments at the cost of a certain loss of information and on the other hand being too permissive retaining more information but keeping also more "noise" brought by unreliable instrumental variables.

In addition, causal effects like the ones we are aiming to discover and highlight, are usually very small, and therefore single studies like ours are often underpowered and could miss many interesting effects due to small sample size.

Nonetheless, some confirmatory findings along with biological plausible pathways make it seem that our is a very promising exploratory approach, and that such an approach could be useful in particular to pinpoint and prioritize risk factors of interest among large sets of candidates, possibly presenting them as reliable objects for further, more specific studies.

Luckily, results of genome-wide association studies are increasingly made publicly available. Harnessing summary-level data, Mendelian randomization analyses then could reach sufficient statistical power to yield more precise causal effect estimates, so additional analysis involving more subjects and genome-wide set of genetic variants would be useful to further investigate the casual mechanisms underlying MS. This could shed even more light on MS pathogenesis and the

biomarkers identified will bear the potential to be used for the diagnosis, for the discrimination among the different forms of the disease, for the monitoring the disease activity and progression and for predicting therapies responses to tailor future therapies for each patient.

Appendix

Antibody ID	IVs (N°)	Antibody ID	IVs (N°)
cp_hpa001834	43	**cyp24a1_hpa022261**	29
c1qa_hpa002350	39	**heatr3_hpa041990**	29
bcl6_hpa004899	37	**il7_hpa019590**	29
trappc2l_hpa041714	36	**ncam2_hpa030900**	29
ilf2_hpa007484	35	**plek_hpa031838**	29
kiaa0494_hpa011224	35	**sh3bgrl3_hpa030848**	29
serpina3_hpa000893	35	**zmynd19_hpa020642**	29
thap6_hpa035767	34	**ankr1_hpa038736**	28
aqp4_hpa014784	33	**c1qc_hpa001471**	28
ogt_hpa030751	33	**copa_hpa028024**	28
serpina4_hpa003607	32	**c9orf46_hpa011144**	27
alpk2_hpa027377	31	**casp6_hpa011337**	27
dsg1_hpa022128	31	**enw1_hpa003407**	27
itih4_hpa001835	31	**mrps15_hpa028100**	27
mobp_hpa035152	31	**taf8_hpa031734**	27
ngfr_hpa004765	30	**triobp_hpa003747**	27
trm13_hpa028494	30	**cndp1_hpa008933**	26
bnip3_hpa003015	29	**ptger4_hpa012756**	26

Antibody ID	IVs (N°)
atp5i_hpa035010	25
casp8_hpa001302	25
clec4a_hpa007842	25
il16_hpa018467	25
rp3412a911_hpa019601	25
rpa3_hpa005708	25
saps2_hpa030656	25
uqcrfs1_hpa041863	25
azgp1_hpa012582	24
casp1_hpa003056	24
csta_hpa001031	24
fxl18_hpa036049	24
lphn1_hpa037974	24
mansc1_hpa007956	24
mapk1_hpa005700	24
mrc1_hpa004114	24
nucb1_hpa008176	24
scn7a_hpa004879	24
sh2b3_hpa005483	24

Antibody ID	IVs (N°)
snap23_hpa001214	24
anxa1_hpa011271	23
bud13_hpa038341	23
elmo1_hpa017941	23
fbln1_hpa001612	23
gap43_hpa013392	23
hisabp_hisabp	23
ict1_hpa003634	23
ifit5_hpa037957	23
il17re_hpa019011	23
il22ra2_hpa030582	23
map3k14_hpa027269	23
osm_hpa029814	23
rpesp_hpa029595	23
sms_hpa029852	23
vps11_hpa039019	23
zbtb46_hpa013997	23
actr6_hpa038587	22
al1604712_hpa028612	22

Antibody ID	IVs (N°)
c16orf75_hpa040995	22
c3_hpa003563	22
c9orf150_hpa024407	22
cryab_hpa028724	22
ctgf_hpa031075	22
dars_hpa024079	22
ddah1_hpa006308	22
dlst_hpa003010	22
dpm3_hpa014667	22
hspa4_hpa010023	22
kiaa1618_hpa003347	22
kiaa1618_hpa026790	22
rpain_hpa031526	22
samc_hpa026887	22
sertad2_hpa019021	22
snap29_hpa031823	22
tmed9_hpa014650	22
vim_hpa001762	22
zfp36l1_hpa001301	22

Antibody ID	IVs (N°)
atp5j_hpa031069	21
cd163l1_hpa015663	21
cfb04_hpa001817	21
crtac1_hpa008175	21
gas6_hpa008275	21
gper_hpa027052	21
mx1_hpa030917	21
pcp4_hpa005792	21
plau_hpa008719	21
rars_hpa003979	21
ren_hpa005131	21
serpina3_hpa002560	21
tjp1_hpa001636	21
wdr74_hpa037795	21
ace2_hpa000288	20
acsl4_hpa005552	20
agap2mettl1_hpa023474	20
app_hpa001462	20
arg2_hpa000663	20

Antibody ID	IVs (N°)
arhgef3_hpa034715	20
bri_hpa029292	20
c9orf103_hpa020378	20
col15a1_hpa017915	20
fcrl3_hpa015508	20
hif1a_hpa001275	20
map3k14_hpa027270	20
nfkbiz_hpa010547	20
snx2_hpa037400	20
stx11_hpa007992	20
taldo1_hpa040373	20
tas2r60_hpa030416	20
tf_hpa001527	20
tpd52_hpa028427	20
alpk2_hpa029801	19
btn3a1_hpa012565	19
casp10_hpa017059	19
cd226_hpa015715	19
cd99_hpa035304	19

Antibody ID	IVs (N°)
cfb_hpa001832	19
depp_hpa037819	19
dsc2_hpa012615	19
kif5a_hpa004469	19
malt1_hpa003865	19
mtl14_hpa038001	19
ntf3_hpa032000	19
oca2_hpa036403	19
polr3gl_hpa027288	19
pphln1_hpa038902	19
rabbitigg_rigg	19
s100a9_hpa004193	19
slc25a39_hpa026785	19
sod1_hpa001401	19
spp1_hpa005562	19
tjp1_hpa001637	19
tnfsf14_hpa012700	19
wdr12_hpa036389	19
znf821_hpa036372	19

Antibody ID	IVs (N°)
arg1_hpa003595	18
commd3_hpa036584	18
ctss_hpa002988	18
efhc2_hpa034492	18
fbln1_hpa001613	18
il1a_hpa030643	18
irf8_hpa002531	18
lpcat4_hpa030719	18
lrg1_hpa001888	18
morn3_hpa038709	18
ptprz1_hpa015103	18
serpinb2_hpa015480	18
serpine1_hpa001539	18
taf8_hpa031730	18
tgm4_hpa032072	18
agrin_hpa040090	17
ccdc59_hpa038555	17
chch5_hpa038263	17
chd1l_hpa027789	17

Antibody ID	IVs (N°)
dtx3l_hpa010570	17
fadd_hpa001464	17
fibb_hpa001900	17
frs3_hpa030174	17
gc_hpa001526	17
il21_hpa038303	17
irf8_hpa002267	17
kng1_hpa001616	17
mmp19_hpa012845	17
mmp8_hpa021221	17
ndfip1_hpa009682	17
nos2a_hpa003871	17
olig3_hpa018303	17
pdgfa_hpa016613	17
prex1_hpa001927	17
serpina1_hpa000927	17
tjp2_hpa001813	17
tnfsf13_hpa004863	17
triobp_hpa019769	17

Antibody ID	IVs (N°)	Antibody ID	IVs (N°)
ca050_hpa030236	16	ywhab_hpa011212	16
ccdc56_hpa031966	16	al7138907_hpa010967	15
cntn1_hpa041060	16	apex1_hpa002564	15
col6a3_hpa010080	16	azi2_hpa035258	15
dusp8_hpa020071	16	ckb_hpa001254	15
il7r_hpa034514	16	clec16a_hpa035814	15
immt_hpa036164	16	cntf_hpa019654	15
lst2_hpa038175	16	fbln1_hpa001642	15
magg1_hpa030602	16	gimap7_hpa020266	15
map2_hpa008273	16	il4_hpa007714	15
mdh1_hpa027296	16	mapk1_hpa003995	15
mertk_hpa036196	16	nrg1_hpa010964	15
pdgfb_hpa011972	16	park7_hpa004190	15
psmc2_hpa019238	16	rnf141_hpa018133	15
rec8_hpa031729	16	romo1_hpa012782	15
rpain_hpa023924	16	sptan1_hpa007927	15
sertad2_hpa020904	16	ttc1_hpa036557	15
timm10_hpa039946	16	ubash3b_hpa038605	15
xpc_hpa035706	16	zdhhc18_hpa040234	15

Antibody ID	IVs (N°)
acyp2_hpa035301	14
b2m_hpa006361	14
casp3_hpa002643	14
elov7_hpa036337	14
erp29_hpa039363	14
exosc10_hpa028484	14
mitd1_hpa036162	14
mtpn_hpa019735	14
nalcn_hpa031889	14
nrp1_hpa030278	14
ppm1d_hpa022277	14
slc30a7_hpa018034	14
sst_hpa019472	14
tmbim1_hpa012093	14
tmem39a_hpa039140	14
tnfsf14_hpa026919	14
a2m_hpa002265	13
calb1_hpa023099	13
casp8_hpa005688	13

Antibody ID	IVs (N°)
cd14_hpa002127	13
chst12_hpa041680	13
cst3_hpa013143	13
gimap7_hpa020268	13
gnl2_hpa027163	13
kiaa0564_hpa039075	13
smyd2_hpa029023	13
syk_hpa001384	13
tthy_hpa002550	13
ywhag_hpa026918	13
agt_hpa001557	12
ahsg_hpa001524	12
c1orf182_hpa028149	12
cd71_hpa028598	12
chd1l_hpa028670	12
cnpase_hpa023278	12
eomes_hpa028896	12
gda_hpa019352	12
hadhb_hpa037539	12

Antibody ID	IVs (N°)
il23a_hpa001554	12
kiaa0494_hpa011938	12
mmp25_hpa036376	12
mycbp2_hpa039945	12
ppp2r5d_hpa029046	12
ptger4_hpa011226	12
rreb1_hpa001756	12
s100a8_hpa024372	12
serpina1_hpa001292	12
spag16_hpa037542	12
tnks_hpa025690	12
xpa_hpa030997	12
zn343_hpa030587	12
znf740_hpa035691	12
aldh5a1_hpa029715	11
cdkn1c_hpa002924	11
chgb_hpa008759	11
clec16a_hpa035815	11
ctgf_hpa031074	11

Antibody ID	IVs (N°)
dsc2_hpa011911	11
fas_hpa027444	11
fbln2_hpa001934	11
gfi1b_hpa007012	11
icam1_hpa004877	11
klk6_hpa019525	11
lamp2_hpa029100	11
mag_hpa012499	11
plat_hpa003412	11
s100b_hpa015768	11
sorbs2_hpa036754	11
angi_hpa036018	10
c1orf106kif21b_hpa027511	10
cfi_hpa001143	10
edn2_hpa028459	10
mmp9_hpa001238	10
mog_hpa021873	10
pja2_hpa040347	10
prickle4_hpa031240	10

Antibody ID	IVs (N°)
sec14l1_hpa028703	10
symm_hpa035590	10
tagap_hpa031000	10
ablm2_hpa035808	9
cflar_hpa019044	9
dpf2_hpa020880	9
eif3h_hpa023117	9
grm7_hpa036659	9
gstk1_hpa022904	9
hpgd_hpa005679	9
igfl1_hpa014001	9
il22_hpa023684	9
kcnrg_hpa001741	9
lck_hpa003494	9
mrps22_hpa006083	9
nlk_hpa018192	9
nrcam_hpa012606	9
slc12a5cd40_hpa004942	9
sorbs1_hpa036994	9

Antibody ID	IVs (N°)
tfpt_hpa034958	9
zbtb16_hpa001499	9
alpk2_hpa027976	8
c9_hpa029577	8
cf081_hpa030894	8
col15a1_hpa017913	8
dop1_hpa027904	8
grp78_hpa038845	8
hladqb1_hpa013667	8
hyls1_hpa041210	8
ifit5_hpa037958	8
igfl1_hpa014270	8
lta_hpa007729	8
mmp8_hpa022935	8
omg_hpa012693	8
rab11fip1_hpa024010	8
rec8_hpa031727	8
taf8_hpa031731	8
ywhae_hpa008445	8

Antibody ID	IVs (N°)	Antibody ID	IVs (N°)
ahsg_hpa001525	7	grm7_hpa015964	6
cd14_hpa001887	7	mapk1_hpa030069	6
crp_hpa027396	7	metap2_hpa019095	6
evi5_hpa027339	7	mki67_hpa000451	6
il18_hpa003980	7	mmel1_hpa008205	6
pib5painpp5j_hpa034539	7	mpv17l2_hpa043111	6
s11ip_hpa036837	7	znf438_hpa039843	6
s19a1_hpa038117	7	c1orf182_hpa029897	5
sfn_hpa011105	7	ccl2_hpa019163	5
sp140_hpa006162	7	fund1_hpa038773	5
stat4_hpa001860	7	wdr91_hpa031520	5
stat6_hpa001861	7	cnpase_hpa023280	4
ttc17_hpa038508	7	lif_hpa018844	4
zn300_hpa028975	7	plg_hpa021602	4
cnpase_hpa023338	6	casp2_hpa006704	3
dffa_hpa019938	6	il12a_hpa001886	3
gda_hpa024099	6	socs6_hpa035477	3

TABLE 15: Number of selected Instrumental Variables (IVs) for each protein (antibody ID shown) analysed, ordered higher to lower.

References

1. Smith, G.D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1-22 (2003).

2. Hingorani, A. & Humphries, S. Nature's randomised trials. *Lancet* **366**, 1906-1908 (2005).

3. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

4. Swerdlow, D.I.*, et al.* Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology* **45**, 1600-1616 (2016).

5. Paternoster, L., Tilling, K. & Davey Smith, G. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. *PLOS Genetics* **13**, e1006944 (2017).

6. Reich, D.S., Lucchinetti, C.F. & Calabresi, P.A. Multiple Sclerosis. *New England Journal of Medicine* **378**, 169-180 (2018).

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

8. Noseworthy , J.H., Lucchinetti , C., Rodriguez , M. & Weinshenker , B.G. Multiple Sclerosis. *New England Journal of Medicine* **343**, 938-952 (2000).

9. Wu, G.F. & Alvarez, E. The immunopathophysiology of multiple sclerosis. *Neurol Clin* **29**, 257-278 (2011).

10. Hemmer, B., Kieseier, B., Cepok, S. & Hartung, H.P. New immunopathologic insights into multiple sclerosis. *Curr Neurol Neurosci Rep* **3**, 246-255 (2003).

11. Wootla, B., Eriguchi, M. & Rodriguez, M. Is Multiple Sclerosis an Autoimmune Disease? *Autoimmune Diseases* **2012**, 969657 (2012).

12. Kenealy, S.J., Pericak-Vance, M.A. & Haines, J.L. The genetic epidemiology of multiple sclerosis. *J Neuroimmunol* **143**, 7-12 (2003).

13. Kurtzke, J.F. An evaluation of the geographic distribution of multiple sclerosis. *Acta Neurol Scand* **42**, Suppl 19:91+ (1966).

14. Ranzato, F.*, et al.* Increasing frequency of multiple sclerosis in Padova, Italy: a 30 year epidemiological survey. *Mult Scler* **9**, 387-392 (2003).

15. Montomoli, C.*, et al.* Multiple sclerosis recurrence risk for siblings in an isolated population of Central Sardinia, Italy. *Genet Epidemiol* **22**, 265-271 (2002).

16. Compston, D.A., Batchelor, J.R. & McDonald, W.I. B-lymphocyte alloantigens associated with multiple sclerosis. *Lancet* **2**, 1261-1265 (1976).

17. Terasaki, P.I., Park, M.S., Opelz, G. & Ting, A. Multiple sclerosis and high incidence of a B lymphocyte antigen. *Science* **193**, 1245 (1976).

18. McKelvey, C. *159 Genetic Variants Now Known to Be Associated With MS* (2014).

19. 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).

20. O'Gorman, C., Lucas, R. & Taylor, B. Environmental Risk Factors for Multiple Sclerosis: A Review with a Focus on Molecular Mechanisms. in *Int J Mol Sci* 11718-11752 (2012).

21. Lev, N.*, et al.* Experimental encephalomyelitis induces changes in DJ-1: implications for oxidative stress in multiple sclerosis. *Antioxid Redox Signal* **8**, 1987-1995 (2006).

22. Milo, R. & Kahana, E. Multiple sclerosis: geoepidemiology, genetics and the environment. *Autoimmun Rev* **9**, A387-394 (2010).

23. Chang, A.*, et al.* Neurogenesis in the chronic lesions of multiple sclerosis. in *Brain* 2366-2375 (2008).

24. Poser, C.M. & Brinar, V.V. Diagnostic criteria for multiple sclerosis. *Clin Neurol Neurosurg* **103**, 1-11 (2001).

25. Cottrell, D.A.*, et al.* The natural history of multiple sclerosis:a geographically based study5. The clinical features and natural history of primary progressive multiple sclerosis. **122**, 625-639 (1999).

26. 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).

27. Hirotani, M.*, et al.* Correlation between DJ-1 levels in the cerebrospinal fluid and the progression of disabilities in multiple sclerosis patients. *Multiple Sclerosis Journal* **14**, 1056-1060 (2008).

28. Compston, A. & Coles, A. Multiple sclerosis. *Lancet* **372**, 1502-1517 (2008).

29. Contu, D.*, et al.* Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans. *PLoS One* **3**, e1430 (2008).

30. Granieri, E.*, et al.* The increasing incidence and prevalence of MS in a Sardinian province. *Neurology* **55**, 842-848 (2000).

31. Varilo, T. & Peltonen, L. Isolates and their potential use in complex gene mapping efforts. *Curr Opin Genet Dev* **14**, 316-323 (2004).

32. Arcos-Burgos, M. & Muenke, M. Genetics of population isolates. *Clin Genet* **61**, 233-247 (2002).

33. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. *Nat Rev Genet* **1**, 182-190 (2000).

34. Bielekova, B. & Martin, R. Development of biomarkers in multiple sclerosis. *Brain* **127**, 1463-1478 (2004).

35. Pearl, J. Causal inference in statistics: An overview. **3**, 96-146 (2009).

36. Hernán, M.A. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *American Journal of Public Health* **108**, 616-619 (2018).

37. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference* (Morgan Kaufmann Publishers Inc., 1988).

38. Hulswit, M. *A Short History of Causation* (2004).

39. A. Fisher, R. *The Arrangement of Field Experiments* (1925).

40. Fisher, R.A. *The Design of Experiments* (Oliver and Boyd, 1935).

41. Yates, F. *The Design and Analysis of Factorial Experiments* (Imperial Bureau of Soil Science, 1937).

42. Poincaré, I.H. *Annales de l'Institut Henri Poincaré: recueil de conférences et mémoires de calcul des probabilités et physique théorique* (Institut Henri Poincaré., 1951).

43. Box, J.F. R. A. Fisher and the Design of Experiments, 1922-1926. *The American Statistician* **34**, 1-7 (1980).

44. Shipley, B. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference* (2000).

45. Cochran, W.G.C.F.p.d.S. Analysis of Covariance: Its Nature and Uses. *Biometrics* **13**, 261-281 (1957).

46. Cox, D.R. Regression Analysis when there is Prior Information about Supplementary Variables. *Journal of the Royal Statistical Society. Series B (Methodological)* **22**, 172-176 (1960).

47. Cox, D.R. & McCullagh, P. Some aspects of analysis of covariance. *Biometrics* **38**, 541-561 (1982).

48. Cochran, W.G. & Chambers, S.P.C.F.p.d. The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society. Series A (General)* **128**, 234-266 (1965).

49. Hill, A.B. The Environment and Disease: Association or Causation? *Proc R Soc Med* **58**, 295-300 (1965).

50. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688-701 (1974).

51. Splawa-Neyman, J. *On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated from the Polish and edited by D. M. Dąbrowska and T. P. Speed* (1990).

52. Dawid, A.P. Causal Inference Without Counterfactuals. *Journal of the American Statistical Association* **95**, 407-424 (2000).

53. M. Robins, J. *Causal Inference from Complex Longitudinal Data* (1999).

54. van der Laan, M.J. & Robins, J.M. *Unified Methods for Censored Longitudinal Data and Causality* (Springer New York, 2012).

55. Rosenbaum, P.R. *Observational Studies* (Springer, 2002).

56. Cox, D.R. & Wermuth, N. Causality: a Statistical View. *International Statistical Review* **72**, 285-305 (2004).

57. Berzuini, C., Dawid, P. & Bernardinell, L. *Causality: Statistical Perspectives and Applications* (Wiley, 2012).

58. Dawid, A.P. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)* **41**, 1-31 (1979).

59. Greenland, S., Pearl, J. & Robins, J.M. Causal Diagrams for Epidemiologic Research. *Epidemiology* **10**, 37-48 (1999).

60. Greenland, S. & Pearl, J. Causal Diagrams. *Wiley StatsRef: Statistics Reference Online* (2017).

61. Fletcher, R.H. & Fletcher, S.W. *Clinical Epidemiology: The Essentials* (Lippincott Williams & Wilkins, 2005).

62. George, E.P.B.C.F.p.d.N. Use and Abuse of Regression. *Technometrics* **8**, 625-629 (1966).

63. Rubin, D.B. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics* **6**, 34-58 (1978).

64. Rosenbaum, P.R. From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment. *Journal of the American Statistical Association* **79**, 41-48 (1984).

65. Burgess, S. & Thompson, S.G. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation* (CRC Press, 2015).

66. Wehby, G.L., Ohsfeldt, R.L. & Murray, J.C. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Stat Med* **27**, 2745-2749 (2008).

67. Thomas, D.C. & Conti, D.V. Commentary: The concept of 'Mendelian Randomization'. **33**, 21-25 (2004).

68. Davey Smith, G. Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. *Genes Nutr* **6**, 27-43 (2011).

69. Nitsch, D*., et al.* Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. in *Am J Epidemiol* 397-403 (United States, 2006).

70. Didelez, V. & Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. in *Stat Methods Med Res* 309-330 (England, 2007).

71. Dawid, A.P. Influence Diagrams for Causal Modelling and Inference. *International Statistical Review / Revue Internationale de Statistique* **70**, 161-189 (2002).

72. Fazia, T*., et al.* Investigating multiple sclerosis genetic susceptibility on the founder population of east-central Sardinia via association and linkage analysis of immune-related loci. *Mult Scler*, 1352458517732841 (2017).

73. Nilsson, P*., et al.* Towards a human proteome atlas: high-throughput generation of mono-specific antibodies for tissue profiling. *Proteomics* **5**, 4327-4337 (2005).

74. Schwenk, J.M., Lindberg, J., Sundberg, M., Uhlen, M. & Nilsson, P. Determination of binding specificities in highly multiplexed bead-based assays for antibody proteomics. *Mol Cell Proteomics* **6**, 125-132 (2007).

75. Schwenk, J.M., Gry, M., Rimini, R., Uhlen, M. & Nilsson, P. Antibody suspension bead arrays within serum proteomics. *J Proteome Res* **7**, 3168-3179 (2008).

76. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* **78**, 4281-4290 (2006).

77. Burgess, S., Small, D.S. & Thompson, S.G. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res* **26**, 2333-2355 (2017).

78. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89-98 (2014).

79. Burgess, S., Butterworth, A. & Thompson, S.G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37**, 658-665 (2013).

80. Pinheiro, J.C. & Bates, D.M. *Mixed-Effects*

Models in S and S-PLUS (2000).

81. Chen, M.H. & Yang, Q. GWAF: an R package for genome-wide association analyses with family data. *Bioinformatics* **26**, 580-581 (2010).

82. Liang, K.-Y. & Zeger, S.L. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **73**, 13-22 (1986).

83. Højsgaard, S., Yan, J. & Halekoh, U. *The R package GEEPACK for generalized estimating equations* (2005).

84. Burgess, S., Dudbridge, F. & Thompson, S.G. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat Med* **35**, 1880-1906 (2016).

85. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-525 (2015).

86. Bowden, J., Davey Smith, G., Haycock, P.C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304-314 (2016).

87. Yavorska, O.O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* (2017).

88. Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**, 1133-1163 (2008).

89. Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**, 629-634 (1997).

90. Kolesár, M., Chetty, R., Friedman, J., Glaeser, E. & Imbens, G.W. Identification and Inference With Many Invalid Instruments. *Journal of Business & Economic Statistics* **33**, 474-484 (2015).

91. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).

92. Welsh, P.*, et al.* Unraveling the directional link between adiposity and inflammation: a bidirectional Mendelian randomization approach. *J Clin Endocrinol Metab* **95**, 93-99 (2010).

93. Mack, S., Coassin, S., Vaucher, J., Kronenberg, F. & Lamina, C. Evaluating the Causal Relation of ApoA-IV with Disease-Related Traits - A Bidirectional Two-sample Mendelian Randomization Study. *Sci Rep* **7**, 8734 (2017).

94. Kemp, J.P., Sayers, A., Smith, G.D., Tobias, J.H. & Evans, D.M. Using Mendelian randomization to investigate a possible causal relationship between adiposity and increased bone mineral density at different skeletal sites in children. *Int J Epidemiol* **45**, 1560-1572 (2016).

95. Kaushansky, N., Eisenstein, M., Zilkha-Falb, R. & Ben-Nun, A. The myelin-associated oligodendrocytic basic protein (MOBP) as a relevant primary target autoantigen in multiple sclerosis. *Autoimmunity Reviews* **9**, 233-236 (2010).

96. de Rosbo, N.K.*, et al.* The Myelin-Associated Oligodendrocytic Basic Protein Region MOBP15–36 Encompasses the Immunodominant Major Encephalitogenic

Epitope(s) for SJL/J Mice and Predicted Epitope(s) for Multiple Sclerosis-Associated HLA-DRB1*1501. *The Journal of Immunology* **173**, 1426 (2004).

97. Holz, A., Bielekova, B., Martin, R. & Oldstone, M.B. Myelin-associated oligodendrocytic basic protein: identification of an encephalitogenic epitope and association with multiple sclerosis. *J Immunol* **164**, 1103-1109 (2000).

98. Simon, R.P., Griffiths, T., Evans, M.C., Swan, J.H. & Meldrum, B.S. Calcium overload in selectively vulnerable neurons of the hippocampus during and after ischemia: an electron microscopy study in the rat. *J Cereb Blood Flow Metab* **4**, 350-361 (1984).

99. Marambaud, P., Dreses-Werringloer, U. & Vingtdeux, V. Calcium signaling in neurodegeneration. *Mol Neurodegener* **4**, 20 (2009).

100. Cataldi, M. The Changing Landscape of Voltage-Gated Calcium Channels in Neurovascular Disorders and in Neurodegenerative Diseases. in *Curr Neuropharmacol* 276-297 (2013).

101. Hoffmann, D.B.*, et al.* Calcium influx and calpain activation mediate preclinical retinal neurodegeneration in autoimmune optic neuritis. *J Neuropathol Exp Neurol* **72**, 745-757 (2013).

102. Barnham, K.J. & Bush, A.I. Metals in Alzheimer's and Parkinson's Diseases. *Current Opinion in Chemical Biology* **12**, 222-228 (2008).

103. Jellinger, K.A. Chapter One - The Relevance of Metals in the Pathophysiology of Neurodegeneration, Pathological Considerations. in *International Review of Neurobiology* (ed. K.P. Bhatia & S.A. Schneider) 1-47 (Academic Press, 2013).

104. Benn, M. & Nordestgaard, B.G. From genome-wide association studies to Mendelian randomization: novel opportunities for understanding cardiovascular disease causality, pathogenesis, prevention, and treatment. cvy045-cvy045 (2018).

105. Conde, S.*, et al.* Mendelian randomisation analysis of clustered causal effects of body mass on cardiometabolic biomarkers. *BMC Bioinformatics* **19**, 195 (2018).

106. Mokry, L.E.*, et al.* Obesity and Multiple Sclerosis: A Mendelian Randomization Study. *PLOS Medicine* **13**, e1002053 (2016).

107. Aikens, R.C.*, et al.* Systolic Blood Pressure and Risk of Type 2 Diabetes: A Mendelian Randomization Study. *Diabetes* **66**, 543-550 (2017).