Dipartimento di scienze economiche, aziendali, matematiche e statistiche "Bruno de Finetti"

Research Paper Series, N. 1, 2019

Bayesian Mendelian Randomization for incomplete pedigree data, and the characterisation of Multiple Sclerosis proteins

TERESA FAZIA
Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy

LEONARDO EGIDI
Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti" (DEAMS), University of Trieste, Trieste, Italy

BURCU AYOGLU
SciLifeLab, Dept. of Protein Science, KTH Royal Institute of Technology, Stockholm, Sweden

ASHLEY BEECHAM
John P. Hussmann Institute for Human Genomics and Dr John Macdonald Foundation, Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, USA

PIER PAOLO BITTI
Azienda Tutela Salute Sardegna. ASSL Nuoro. Immunoematologia e Medicina Trasfusionale. Ospedale "San Francesco", Nuoro, Italy

ANNA TICCA
Azienda Tutela Salute Sardegna. ASSL Nuoro. Neurologia e Stroke Unit. Ospedale "San Francesco", Nuoro, Italy

JACOB L. McCAULEY
John P. Hussmann Institute for Human Genomics and Dr John Macdonald Foundation, Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, USA

PETER NILSSON
SciLifeLab, Dept. of Protein Science, KTH Royal Institute of Technology, Stockholm, Sweden

CARLO BERZUINI
Centre for Biostatistics, The University of Manchester, Jean McFarlane Building, University Place, Oxford Road, Manchester M13 9PL, UK

LUISA BERNARDINELLI
Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Research Paper Series

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche
"Bruno de Finetti"
Via dell'Università 1
34123, Trieste
Tel. +390405587927
Fax +390405587005
http://www.deams.units.it

EUT Edizioni Università di Trieste
Via E.Weiss, 21 - 34128 Trieste
Tel. +390405586183
Fax +390405586185
http://eut.units.it
eut@units.it

ISBN: 978-88-5511-114-0

Research Paper Series, N. 1, 2019

Bayesian Mendelian Randomization for incomplete pedigree data, and the characterisation of Multiple Sclerosis proteins

TERESA FAZIA
Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy

LEONARDO EGIDI
Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti" (DEAMS), University of Trieste, Trieste, Italy

BURCU AYOGLU
SciLifeLab, Dept. of Protein Science, KTH Royal Institute of Technology, Stockholm, Sweden

ASHLEY BEECHAM
John P. Hussmann Institute for Human Genomics and Dr John Macdonald Foundation, Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, USA

PIER PAOLO BITTI
Azienda Tutela Salute Sardegna. ASSL Nuoro. Immunoematologia e Medicina Trasfusionale. Ospedale "San Francesco", Nuoro, Italy

ANNA TICCA
Azienda Tutela Salute Sardegna. ASSL Nuoro. Neurologia e Stroke Unit. Ospedale "San Francesco", Nuoro, Italy

JACOB L. McCAULEY
John P. Hussmann Institute for Human Genomics and Dr John Macdonald Foundation, Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, USA

PETER NILSSON
SciLifeLab, Dept. of Protein Science, KTH Royal Institute of Technology, Stockholm, Sweden

CARLO BERZUINI
Centre for Biostatistics, The University of Manchester, Jean McFarlane Building, University Place, Oxford Road, Manchester M13 9PL, UK

LUISA BERNARDINELLI
Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy

ABSTRACT[1]

Before the GWAS (genome-wide association study) era, many genetic determinants of disease were found via analysis of multiplex pedigrees, that is, by looking for genetic markers that run in families in a similar way as disease. GWAS advent has robbed pedigree analysis of its luster. Future scientific methodology seesaw might bring pedigree analysis back into the spotlight. After the recent discovery of hundreds of disease-associated variants, interest is focusing on the way these variants affect downstream molecular markers, such as transcripts and protein levels, and on the way the resulting changes in these markers in turn affect disease risk. Statistical methods such as Mendelian Randomization (Katan, 1986), hereafter denoted as MR, represent important tools in this effort. Most MR studies are based on data from unrelated individuals, a notable exception being Brumpton et al. (2019). In the present paper we argue that by enriching these data with data from family-related individuals, a number of difficulties that are encountered in MR can be significantly attenuated. Motivated by the above considerations, this paper discusses extensions of MR to deal with pedigree data. We adopt the Bayesian MR framework proposed by Berzuini and colleagues (Berzuini et al., 2018), and extend it in various ways to deal with pedigree data.

KEYWORDS: Genetic determinants, Pedigree, Mendelian Randomization, Bayesian methods

[1]**Corresponding author:** Leonardo Egidi, Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti", Via dell'Università 1, 34123, Trieste, Italia, email: legidi@units.it tel: 0405587047; fax: 040.5587005

1. Introduction

Before the GWAS (genome-wide association study) era, many genetic determinants of disease were found via analysis of multiplex pedigrees, that is, by looking for genetic markers that run in families in a similar way as disease. GWAS advent has robbed pedigree analysis of its luster. Future scientific methodology seesaw might bring pedigree analysis back into the spotlight.

After the recent discovery of hundreds of disease-associated variants, interest is focusing on the way these variants affect downstream molecular markers, such as transcripts and protein levels, and on the way the resulting changes in these markers in turn affect disease risk. Statistical methods such as Mendelian Randomization (Katan, 1986), hereafter denoted as MR, represent important tools in this effort. Most MR studies are based on data from unrelated individuals, a notable exception being Brumpton et al. (2019). In the present paper we argue that by enriching these data with data from family-related individuals, a number of difficulties that are encountered in MR can be significantly attenuated.

Motivated by the above considerations, this paper discusses extensions of MR to deal with pedigree data. We adopt the Bayesian MR framework proposed by Berzuini and colleagues (Berzuini et al., 2018), and extend it in various ways to deal with pedigree data. The proposed method exploits recent developments in Markov chain Monte Carlo (MCMC) inference, as offered by the `Stan` probabilistic programming language (Carpenter et al., 2017).

We illustrate the method with the aid of data generated by ImmunoChip genotyping and transcriptome/protein assays on members of Multiple Sclerosis (MS) multiplex pedigrees from an isolated Sardinian (italian island) population. With this kind of data, environmental confounding and population stratification are expected to have less impact on causal effect estimates, and the effects of rare variants to be easier to detect. Thanks to our Bayesian technology, we perform a "clever" analysis where an initial model is gradually elaborated to bring biological theory and relevant information in general to bear. In this paper, we include in the MR model such information as a family indicator, parental protein levels and kinship. Not only do such enhancements provide extra protection against bias, but they also allow us to explore a number of secondary aspects of the biological mechanism. A further advantage of the Bayesian approach is the simple way it deals with incomplete information. In our study, missing values of the exposure (the level of a protein) are treated as additional parameters to be estimated from the data, without incurring biases, as is natural in Bayesian analysis.

The "outcome" variable of our analysis is the MS disease indicator. MS lends itself well to a MR study. This disease tends to become manifest early during reproductive lifespan of most humans, throughout history, and is therefore likely to have a strong genetic component. Genetic variants are therefore expected to act as good instruments for the MR analysis. The main scientific question in this paper is whether the plasma level of IL12A protein (which in our analysis will be referred to as the "exposure") is causal with respect to development of MS (outcome). It is believed that dysregulation of circulating proteins is a causal determinant in many pathologies, more directly so than genetic variants. Our analysis is further motivated by the importance of proteins as

natural drug targets. We could have harnessed publicly available eQTL information to involve in the analysis protein concentrations in tissues other than blood, but we do not pursue this here, not to obscure the main points of the paper, whose main message is methodological.

2. Methods

a. Sample Description

Our MS patients were ascertained through the case register established in 1995 in the province of Nuoro, Sardinia, Italy. Cases were diagnosed according to Poser's criteria (Poser et al., 1983). Twenty extended MS multiplex pedigrees were selected for the analysis, for a total of $N = 936$ individuals (98 cases and 838 unaffected relatives). A subset of the pedigree members had complete data, consisting of the observed levels of the IL12A protein (the exposure), the known disease indicator (the outcome variable), and the genotypes at all loci of Immunochip (see below). The remaining individuals had complete data except for a missing value for the protein level.

b. Genotyping Data

Genotyping data were obtained by using Immunochip Illumina Infinium HD custom array (hereafter "Immunochip" for brevity), designed for fine mapping of 184 established autoimmune loci (Beecham et al., 2013).

The quality control-filtered dataset included 127134 Single Nucleotide Polymorphisms (SNPs) across Immunochip (Fazia et al., 2017). For a first stage of our analysis, we imposed a maximum correlation of $r^2 = 0.20$ between candidate instrumental SNPs within a 100 Kb window, by using the `indep-pairwise` command of the PLINK package (Pur, 2007). This yielded a total of 19121 candidate SNP instruments across Immunochip.

c. Protein Selection and Profiling

The protein we chose for our illustrative study was IL12A. Choice was made prior to considering the data, on the basis of Genome-Wide Significant (GWS) association between MS and genetic variants located within (e.g. exonic, intronic, in the UTR) or in the proximity (e.g. downstream, intergenic) of the protein-coding gene (Beecham et al., 2013) and on the basis of literature evidence on the biological role of this cytokine in the context of MS (Constantinescu et al., 1998; Jahanbani-Ardakani et al., 2019; Rentzos et al., 2008; Sun et al., 2015). Detailed information about the locations of the strongest MS association signals within or in the proximity of the protein-coding genes, and about the strengths of the MS associations, are reported for IL12A in the Supplementary Material.

Plasma profiles were analysed by using a bead-based antibody array format, consisting of polyclonal Human Protein Atlas (Nilsson et al., 2005) antibodies immobilized onto microspheres in suspension (Schwenk et al., 2007, 2008) (see Supplementary Material for details).

d. Selection of Instrumental Variants

Genetic variants with a significant marginal association ($p < 5 \times 10^{-3}$) with the level of the protein of interest and mutual $r^2 < 0.20$ correlation were selected to act as instrumental variables (IVs) in the first stage of our analysis. The liberal $p < 5 \times 10^{-3}$ threshold is justified by the fine genotyping of candidate gene regions and by recent arguments (X et al., 2016; J et al., 2011) in favour of using sub-genome-wide-significance loci to strengthen biologically interesting signals. It is also justified by the relative ability of our Bayesian MR method (when compared with most frequentist approaches) to deal with the weak instrument bias, thanks to the uncertainty of the estimated exposure coefficients being explicitly included in the model.

e. Notation

In our analysis, the putative causal factor (with respect to disease) is the circulating level of protein IL12A. We call this variable the "exposure", and denote it as X. We let the symbol Σ_X denote a regime indicator (Dawid, 2000, 2002) which tells us whether we are considering the *actual* data generating regime for X, which is observational, or a *hypothetical* regime where variable X in each individual is set to a value x by intervention. The observational regime corresponds to $\Sigma_X = \emptyset$, whereas the latter, interventional, regime corresponds to $\Sigma_X = x$. In our analysis the outcome variable, Y, indicates whether the individual has the disease ($Y = 1$) or not ($Y = 0$). We are interested in the "causal effect" of X on Y, that is, in the way the distribution of Y changes when X is first set by intervention to a reference value x_0 and then forced to take the new value x_1. Throughout this paper we take this causal effect to be defined as the causal odds ratio (COR):

$$COR = \frac{P(Y = 1 \mid \Sigma_X = x_1)}{P(Y = 1 \mid \Sigma_X = x_0)} \frac{1 - P(Y = 1 \mid \Sigma_X = x_0)}{1 - P(Y = 1 \mid \Sigma_X = x_1)} \tag{1}$$

The reason why we can't generally measure causal effect by standard regression of Y on X is that the regression coefficient will have no causal interpretation in the presence of unobserved confounders of the exposure-outcome relationship, which we denote as U. This is, indeed, why we need to use MR. We shall model U as an individual-level scalar variable, more precisely, a one-dimensional reduction of the unknown collection of confounders. MR requires availability of a set of instrumental variables, or instruments, denoted as $Z \equiv (Z_1, \ldots, Z_J)$, which in a standard analysis will often correspond to the individual's genotypes at a set of SNP loci. Each of these genotypes we code as "allele doses", with values $(0, 1, 2)$ respectively indicating presence of zero, one and two copies of the "alternative" allele at the locus. For most individuals in the pedigree, we also have observed *(i)* maternal and paternal genotypes at each instrumental locus and *(ii)* the levels of protein IL12A in the father and in the mother. Let the collection of maternal (rep., paternal) genotypes for the generic individual be denoted as Z_m (Z_p). Let the protein levels for the mother and the father of the generic individual be denoted as $W_M$and W_F, respectively. We further introduce an individual-level categorical variable, denoted as F, which indicates the individual's pedigree of membership, or family. Further notation will be introduced in the next sections, as required.

f. Assumptions

This paper uses Dawid's conditional independence formalism (Dawid, 1979), with the $\perp\!\!\!\perp$ symbol representing conditional independence, so that $A \perp\!\!\!\perp B \mid C$, stands for "$A$ is independent of B given C, and $A \not\perp\!\!\!\perp B$, means "A is not independent of B ". Conditions introduced in this section are required for method validity, except for one of them. They are essentially identical to those required by standard MR methods.

Here are the assumptions. Each jth instrumental variable, Z_j, must satisfy the *confounder independence* condition $Z_j \perp\!\!\!\perp U$, stating that the instrument is unrelated to exposure-outcome confounders. A further condition called *exclusion-restriction* requires that $Y \perp\!\!\!\perp Z_j \mid (X, U)$, that is, each jth instrument can be associated with response only via the exposure. Exclusion-restriction is a desirable condition, however, unlike the remaining conditions in this section, it is not required by our method. Next comes the *instrument relevance* condition, $Z_j \not\perp\!\!\!\perp X$, stating that no instrument is independent of the exposure. We have also conditions involving the regime indicator, Σ_X. The *confounder invariance* condition, $U \perp\!\!\!\perp \Sigma_X$, requires that the distribution of the confounders U be the same, whether or not we intervene on X, and regardless of the value imposed on or observed in X. Next comes the *interventional irrelevance* condition $\Sigma_X \perp\!\!\!\perp Z$, requiring that any intervention on X has no consequence on Z, and the *interventional modularity* condition, $\Sigma_X \perp\!\!\!\perp Y \mid (X, U)$, asserting that once we are told the values of X and U, the distribution of Y no longer depends on the way the value of X has arisen, whether observationally or through the intervention of interest.

Those independence relationships that involve the (non-stochastic) regime indicator should be interpreted in the light of the extended conditional independence calculus described by Constantinou et al (Constantinou and Dawid, 2017). The relationships between Σ_X and the remaining variables, as depicted in Figure 1, characterize the influence of X on Y, corresponding to the $X \to Y$ arrow, as causal. The remaining arrows in the graph, eg $Z \to X$, do not necessarily have to be interpreted as causal, which greatly expands method applicability.

How realistic are the above assumptions? This is a crucial question, considering that all the above assumptions, except for instrumental relevance, are at best only indirectly testable, or corroborated on the basis of bological knowledge. Take, for example, the confounder independence assumption. In our application, where the exposure is a low-level biological mark, it may be reasonable to assume that those genetic variants that operate in *cis* with respect to the studied protein, exert no effect on common causal precursors of exposure and outcome other than effects mediated by the exposure. This assumption can be further corroborated by investigations based on eQTL data and on the known linkage disequilibrium (LD) pattern in the DNA region of interest. The assumption of confounder invariance requires more attention than is usually the case. In our application, for example, if the intervention represented by Σ_X consisted of a particular diet, then confounder invariance would be violated, because a diet will hardly modify the level of the protein without altering a constellation of metabolites that act as potential confounders. Interventional irrelevance is defendable in our applicative situation, by using randomization arguments. As concerns interventional modularity, in our study this condition implies, in particular, that a unit increase in X caused by one of the variants in the instrumental

set should exert on Y the same effect as a unit increase in X caused by the intervention of interest. In our application, where the instrumental effects are regulatory and the intervention of interest consists of a pharmacological modification of X, interventional modularity appears to be a defendable assumption.

All the conditions defined above, except for exclusion-restriction, are required by our method.

Sometimes it is possible, and then helpful, to represent the qualitative structure of a statistical model by a directed acyclic graph (Lauritzen, 1996). A stripped-down representation of the class of MR models discussed in the present paper is shown in Figure 1. All the conditions stated above (except for exclusion-restriction) can be read off the graph of Figure 1 by applying d-separation (Geiger et al., 1990) or moralization (Lauritzen, 1996), with the following additional rules: *(i)* faithfulness (Spirtes et al., 2001) of the $Z \to X$ edges (which means assuming that any distribution which follows the model only exhibits independence relations represented by the directed acyclic graph), and *(ii)* assigning a value x to Σ_X implies the simultaneous assignment of the same value to X, and *(iii)* assigning a value x to Σ_X implies that all arrows into X except for $\Sigma_X \to X$ are severed. Because most of the conditions introduced at the beginning of this section are not directly testable on the basis of the data, the Reader should be aware that graphs like the one shown in Figure 1 describe an *assumed*, ultimately uncertified, albeit plausible, state of affairs. We shall assume throughout the paper that the above described conditions, bar exclusion-restriction, are valid.

We conclude this section with a brief discussion of the exclusion-restriction assumption. This assumption (which is not required by our method) does not allow an instrument to exert an effect on Y other than that exerted though the mediating effect of X. In our graph of Figure 1, this condition is violated by the $Z_J \to Y$ arrow. Because of this, the effect of instrument Z_J on Y is said to be "pleiotropic " according to Figure 1. In the context of our application, pleiotropic effects may arise from two broad classes of mechanism. The first is due to the eQTL variants used as instruments being in linkage disequilibrium (LD) with eQTLs of nearby genes. The second is due to the instrumental variant exerting a causal effect on Y through a pathway independent of X. Although the former type of pleiotropy could, in principle, be neutralized by conditioning on the eQTLs in the region, except for the instrumental variants, the latter cannot be directly tested from the data. It would therefore be uncautious to perform MR by using a method that does *not* allow for general types of pleiotropy. Our Bayesian approach deals with the problem by explicitly introducing the unknown pleiotropic effects in the model, and by treating them as unknown parameters to be estimated from the data.

g. Progressive Elaboration of the Model

A "naive" approach consists of analyzing the pedigree data by using the Bayesian MR model proposed by Berzuini and colleagues, as described in Berzuini et al. (2018), as if the individuals were independent. This will, of course, produce biased estimates. We shall use this "independence model" in a preliminary analysis of the data. We shall then step through a sequence of re-analyses of the data based on more elaborated, and more realistic, models, that we describe in the following.

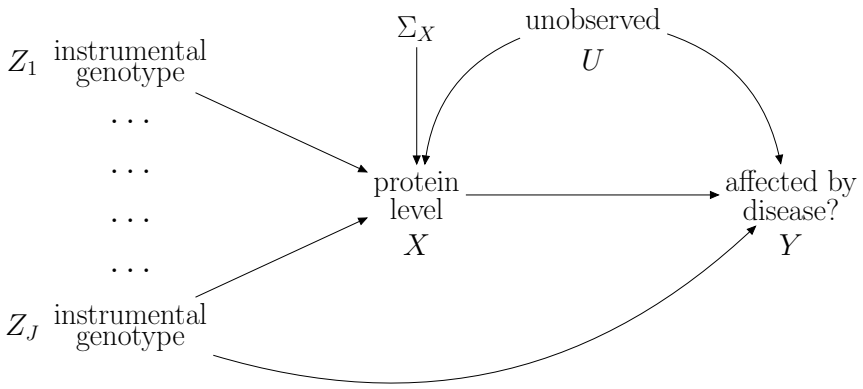

FIG. 1. Graphical representation of a Mendelian randomization model for the analysis of unrelated individuals.

1) INDEPENDENCE MODEL

The model of Berzuini and colleagues (Berzuini et al., 2018) assumes that individuals are independent, and that the X variable has been standardized to have zero mean and unit standard deviation. The data generating equations of the model conform with the conditional independence assumptions expressed in Figure 1, and take the form:

$$P(U) = \mathrm{N}(0,1), \tag{2}$$

$$P(X \mid Z_1,\ldots,Z_J,U) = \mathrm{N}(\sum_{j=1}^{J}\alpha_j Z_j + \delta_X U, \sigma_X^2), \tag{3}$$

$$P(Y \mid X,Z_1,\ldots,Z_J,U) = \mathrm{logit}^{-1}(\omega_Y + \theta X + \sum_{j=1}^{J}\beta_j Z_j + U), \tag{4}$$

where $\mathrm{N}(a,b)$ stands for a normal distribution with mean a and variance b, the symbol $\alpha \equiv (\alpha_1,\ldots,\alpha_J)$ denotes the instrument (i)-exposure (e) associations and $\beta \equiv (\beta_1,\ldots,\beta_J)$ are the pleiotropic effects. The only difference from Berzuini et al here is that the outcome variable Y is no longer normal, but Bernoulli, as appropriate for a binary random variable. Recall that, in our study, some components of the X vector (protein level measurements) are missing, which is not made explicit in the notation. The Bayesian inference engine identifies the missing components and treats them as unknown parameters, effectively integrating them out to obtain the posterior distribution for the parameters of inferential interest. Note that this way of dealing with missing data is more efficient than, say, imputing each missing component of X on the basis of the individual's observed Z values, thanks to the fact that, in our method, the missing values are estimated by using information about both X and Y.

In the above equations, the causal effect of interest, denoted as θ, represents the change in log-odds of probability of $Y = 1$ caused by an interventional change of one standard deviation in X.

As shown in Berzuini et al. (2018) for the normal case, parameters (α, τ_X) are identified by the data, but the remaining parameters, including the causal effect of interest, θ, are not. Berzuini and colleagues deal with the problem by a combination of two devices. The first consists of introducing the additional (untestable) assumption that each jth component of β is a priori independent of the remaining parameters of the model, formally, $P(\beta_j \mid \alpha_j, \tau_X) = P(\beta_j)$. This is called the Instrument Effects Orthogonality (IEO) condition. The second consists of introducing a proper, scientifically plausible, prior for β, which makes inferences possible by inducing on θ (and on further parameters of potential posterior interest) a proper posterior.
As concerns the prior component of our Bayesian model, we invite the Reader to consult (Berzuini et al., 2018).

Variations have been introduced. While still imposing on the pleiotropic effects β a horseshoe prior (Carvalho et al., 2010), we are now using the enhanced version of this distribution proposed by Piironen and Vehtari (Piironen and Vehtari, 2017). Also, we take θ – the causal effect of main inferential interest – to have a Cauchy(0,2.5) prior, with the following justification. Because X has been standardized to have mean 0 and unit standard deviation (SD), the mentioned prior for θ states as unlikely that a one-SD change in protein level causes a change in risk of disease exceeding 5 points on a logit scale, which corresponds to shifting a probability of disease occurrence from, say, 0.01, to 0.5, or from 0.5 to 0.99. This is also in agreement with current evidence on the effect of circulating proteins on disease (Sun et al., 2018).

Finally, we are now taking the i-e associations, α, to be independently distributed according to a double-exponential distribution with mean 0 and unknown scale. One merit of this prior is to shrink the small effects to zero, which reduces the weak instrument bias, so that the model works with an adaptively selected subset of strong instruments.

2) Introducing Kinship

Treating members of a pedigree as independent individuals, which they are not, will produce overconfident and biased estimates. We remedy this by introducing in the model between-individual correlation in the form of the kinship matrix, which can be derived by a standard algorithm from the structure of the pedigree. We are currently working with a single, overarching, kinship matrix of size $N \times N$, where N is the total number of individuals in the sample. This large matrix contains zeros corresponding to pairs of individuals in different families. The method could be made computationally more efficient by introducing family-specific matrices. Kinship information is introduced in the model by writing:

$$
\begin{aligned}
P(Y \mid X, Z_1, \ldots, Z_J, U) &= \text{Bernoulli}(\pi), && (5)\\
logit(\pi) &= \text{MVN}(\mu, \Sigma), && (6)\\
\mu &= \omega_Y + \theta X + \sum_{j=1}^{J} \beta_j Z_j + U, && (7)
\end{aligned}
$$

where Σ is the $N \times N$ kinship matrix, the notation $\text{MVN}(a, b)$ stands for multivariate normal distribution with vector mean a and variance-covariance matrix b.

3) Introducing Family Effects

In our analysis, we incorporate family information simply by designating a categorical variable F to indicate the individual's family, with $F \in (1, \ldots, M)$, with $M = 12$, and by modifying the outcome and exposure models to take the following form:

$$\begin{aligned} P(X \mid Z_1, \ldots, Z_J, U, F) &= \text{N}(\nu, \sigma_X^2), \\ \nu &= \sum_{j=1}^{J} \alpha_j Z_j + \delta_X U + \sum_{f=1}^{M} I_{F=f}\, \gamma_f^X, \\ P(Y \mid X, Z_1, \ldots, Z_J, U, F) &= \text{Bernoulli}(\pi), \\ logit(\pi) &= \text{MVN}(\mu, \Sigma), \\ \mu &= \omega_Y + \theta X + \sum_{j=1}^{J} \beta_j Z_j + U + \sum_{f=1}^{M} I_{F=f}\, \gamma_f^Y, \end{aligned}$$

where I_A stands for the indicator function, taking value 1 if the logical condition A is true, and value 0 otherwise. The quantities $\gamma^X \equiv (\gamma_1^X, \ldots, \gamma_M^X)$ and $\gamma^Y \equiv (\gamma_1^Y, \ldots, \gamma_M^Y)$ are vectors of unknown "family effects", respectively on X and on Y. In our analysis, we have imposed on these parameters independent and mildly informative priors, with greater spread than the prior for θ.

The family indicator appears in the graph of Figure 2 with the symbol F. According to this graph, failure to condition on this indicator (that is, removing the F variable from the model) "opens" (unblocks) the $Z \leftarrow F \rightarrow Y$ path, and the $Z \leftarrow F \rightarrow U$ path, in the terminology of d-separation. Which means that failure to condition on family creates a spurious, exposure-unmediated, association between instrument and outcome and, what's even worse, violates the Confounder Independence assumptions. Hence, inclusion of the family indicator in the model prevents the estimate of the causal effect from being unduly distorted. In situations where the sample contains unrelated (in addition to related) individuals, the unrelateds may be lumped into a single, notional, family.

4) Introducing Parental Protein Information

In this final elaboration step of the model we introduce information about the measured level of protein in the individual's parents. This is motivated by the assumption that there are unobserved loci in DNA, denoted by Z', that (individully or collectively) have an effect on the protein of interest. The individual's protein level becomes associated with that of their parents through Z'. And, because of this, parental protein level become additional candidate instruments in the analysis. We incorporate parental protein information simply by designating the continuous variables P_M and P_F to represent the measured level of circulating IL12A protein in the individual's mother and father, respectively, after standardizing them to have zero mean and unit variance. The two variables are incorporated in the exposure model by writing:

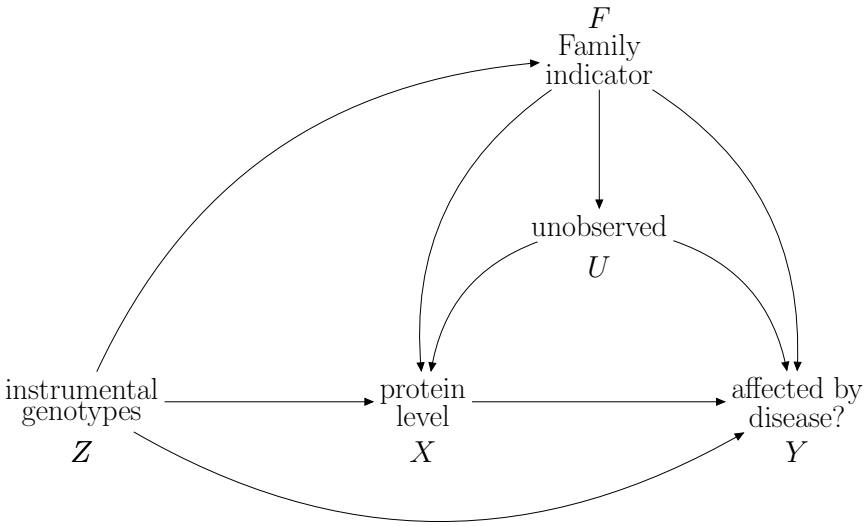

FIG. 2. Incorporating a family indicator variable (F).

$$\begin{aligned} P(X \mid Z_1, \ldots, Z_J, U, F, P_M, P_F) &= \mathrm{N}(\nu, \sigma_X^2), \\ \nu &= \sum_{j=1}^{J} \alpha_j Z_j + \delta_X U + \sum_{f=1}^{M} I_{F=f}\, \gamma_f^X + \alpha^M P_M \alpha^F P_F \end{aligned}$$

with α^M and α^F to be estimated from the data. It can be shown (but this is outside the scope of the present work) that the modification is valid provided we assume that Z and Z' are not correlated, and that Z' does not influence Y other than through changes in X.

3. Results

a. Results from Initial Model

Estimates of the causal effect of the circulating level of IL12A on risk of MS were obtained by using `R` package `MendelianRandomization` (Yavorska and Burgess, 2017), as found on http://cran.r-project.org. The frequentist causal effect estimates, expressed on a log-odds-ratio scale with their corresponding 95% confidence intervals, are summarised in Table 1. Difficulties introduced by the missing IL12A values have been sidestepped in the simplest way: by discarding individuals who had a missing IL12A value when calculating the i-e associations.

According to Table 1, estimates from the frequentist MR methods considered in this paper exhibit a poor consistency. A significant estimate of the causal effect was obtained only with the Simple Median and with the Penalized IVW methods, the latter requiring the assumption of no pleiotropy.

The model by Berzuini and colleagues (Berzuini et al., 2018), which also assumes sample individuals to be independent of each other (see Methods section), gave an estimated

TABLE 1. Estimates for the causal effect of the circulating level of IL12A on risk of MS obtained by using `R` package Mendelian (`http://cran.r-project.org`). Estimated causal effects are expressed on a log-odds-ratio scale.

	Method	Estimate	Std Error	95% conf. interval	P-value
1	Simple median	-0.30	0.15	(-0.59; -0.02)	0.04
2	Weighted median	-0.07	0.14	(-0.34 ; 0.20)	0.61
3	Penalized weighted median	-0.15	0.14	(-0.42 ; 0.12)	0.28
4	IVW	-0.14	0.09	(-0.33 ; 0.04)	0.12
5	Penalized IVW	-0.21	0.10	(-0.40 ; -0.02)	0.03
6	Robust IVW	-0.21	0.12	(-0.44 ; 0.02)	0.08
7	Penalized robust IVW	-0.23	0.10	(-0.42 ; -0.04)	0.02
8	MR-Egger	0.51	0.37	(-0.22 ; 1.25)	0.17
9	Penalized MR-Egger	0.51	0.37	(-0.22 ; 1.25)	0.17
10	Robust MR-Egger	0.51	1.01	(-1.48 ; 2.50)	0.61
11	Penalized robust MR-Egger	0.51	1.01	(-1.48 ; 2.50)	0.61

log-odds-ratio causal effect of -0.202, with a standard error of 0.078, and a 95% credible interval of -0.418 through -0.091. This result was obtained by treating the missing protein levels as additional unknown parameters to be estimated from the data.

b. Results after Introducing Kinship

Our frequentist analyses were repeated in a sounder fashion, by estimating the disease-instrument log-odds-ratio associations via a mixed-effects model (`lmekin` function of `R`, as described in Pinheiro and Bates (2000)), that allows family relationships between pedigree members, as expressed by the kinship matrix, to be taken into account (Fazia et al., 2017). Significant estimates were then obtained by using IVW ($\hat{\theta} = -0.18, p < 0.0001$) and WME ($\hat{\theta} = -0.11, p = 0.012$), but not by using MR-ER ($\hat{\theta} = -0.23, p = 0.7$).

By contrast, when we extended the model by Berzuini and colleagues to incorporate family relationships, as expressed by the kinship matrix (see Methods section), and used it to re-analyse the data, the estimated causal effect was no longer significant, as reported in Table 2. This was not unexpected, when one considers that between-individual correlation reduces the "effective" sample size, and, as a consequence, statistical power.

c. Results after Introducing Family Effects

In the Methods section we have seen that(pedigree) membership may introduce bias in the estimated causal effect by acting as a confounder of the relationship between instrumental genotypes and outcome, in a way similar to what population stratification does. This is a

TABLE 2. Estimates for the causal effect of the circulating level of IL12A on risk of MS obtained by using an extension of the model by Berzuini and colleagues which incorporates family relationships, as expressed by the kinship matrix.

CAUSAL EFFECT OF 1SD CHANGE IN PROTEIN LEVEL ON MS RISK	PERCENTILES OF POSTERIOR DISTRIBUTION				
	5	25	50	75	95
Causal Log Odds Ratio Effect	-0.91	-0.59	-0.39	-0.17	0.10
Causal Odds Ratio Effect	0.4	0.55	0.67	0.84	1.1

TABLE 3. Estimated causal effect of IL12A protein level on MS, expressed on both a log-odds ratio and an odds ratio scale, as obtained by an analysis that incorporates both kinship information and the family indicator.

CAUSAL EFFECT OF 1SD CHANGE IN PROTEIN LEVEL ON MS RISK	PERCENTILES OF POSTERIOR DISTRIBUTION				
	5	25	50	75	95
Causal Exposure Log Odds Ratio	-1.05	-0.69	-0.43	-0.19	0.14
Causal Exposure Odds Ratio	0.35	0.50	0.65	0.82	1.15

consequence of the family variable being generally associated with both the individual's genetic set-up and with disease-linked unobserved factors (genetic variants, environment, education, and so on). See the Methods section for a more rigorous discussion of the issue. When we introduced both kinship information and the family variable (as a 12-level categorical factor) in the model, we got the causal effect estimate summarised in Table 3.

A comparison with the preceding table shows that introduction of the family variable left the point estimate of the causal effect substantially unchanged, while widening the credible interval, with a consequent, further, reduction in statistical significance of the result. This is hardly suprising, when one considers that families 3 and 7 (out of our 12 families) impacted on both exposure and outcome with effects of the same sign, as described later in this section. This will inevitably inflate association between exposure and outcome beyond the component of association due to a genuinely causal relationship.

d. Results from Final Model

In addition to kinship and to the family indicator, our final model includes the measured parental levels of circulating IL12A protein, which means the protein level in the mother and in the father. See Methods section for technical details. This final elaboration increased the amount of instrumental information in the model, and produced the estimates summarized in Table 4. The point estimate for the causal effect of IL12A protein level

TABLE 4. Causal effect estimates from a model that incorporates kinship information, family indicator, and parental protein levels.

CAUSAL EFFECT OF 1SD CHANGE IN PROTEIN LEVEL ON MS RISK	PERCENTILES OF POSTERIOR DISTRIBUTION				
	5	25	50	75	95
Causal Log Odds Ratio Effect of Exposure on Outcome	-1.12	-0.71	-0.49	-0.29	-0.1
Causal Odds Ratio Effect of Exposure on Outcome	0.33	0.49	0.61	0.75	0.90

on risk of MS was -0.49 on a log-odds ratio scale, and 0.61 on an odds-ratio scale. The corresponding 95% credible interval, also reported in Table 4, was entirely contained in the negative real axis, and included effect values of biological importance.

Figures 3 through 5 summarize extra output of the analysis via our final model. These figures have been obtained by using the excellent `bayesplot` package, writen in `R` language by Jonah Gabry and colleagues (Gabry et al., 2019), as an aid to studying the output of `Stan` analyses.

Figure 3 shows posterior intervals of for the instrument-exposure associations, α. It is apparent from the figure that a few instruments, eg. instrument 49, stand out in terms of strength. The sparsity prior we have imposed on these effects is able to pick up the few "needles in the haystack", while downplaying the role of weaker instruments, at the same time working in the direction of a reduction of the weak instrument bias. It might be interesting to investigate the strong instruments from a functional point of view.

Figure 4 shows posterior intervals for familial effects on outcome, that we call "direct" because they are not mediated by the exposure. One may wish to interpret these as familial effects mediated by IL12A-independent pathways, environment and lifestyle. The figure highlights some families (eg., family 12) as characterized by a higher risk of MS, compared with the others. In a separate work we investigate the factors responsible of such differences in detail. Other families (eg., family 2) appear to "protected" from MS due to factors other than IL12A.

Figure 5 shows posterior intervals for familial effects on outcome, that we call "indirect" because they are mediated by the exposure. They are calculated by including in the model a parameter defined to represent the product of the $F \to X$ effect and the $X \to Y$ effect. The posterior distribution for this parameter gets sampled by the MCMC inference engine. The sample are then automatically used to calculate posterior mean and credible interval. A comparison between Figures 4 and 5 suggests that in certain families, eg. family 4 in our sample, both the direct and the indirect effects operate deleteriously, whereas in others, eg. family 7, the two effects tend to cancel each other.

We calculated posterior predictive check diagnostics based on discrepancies between

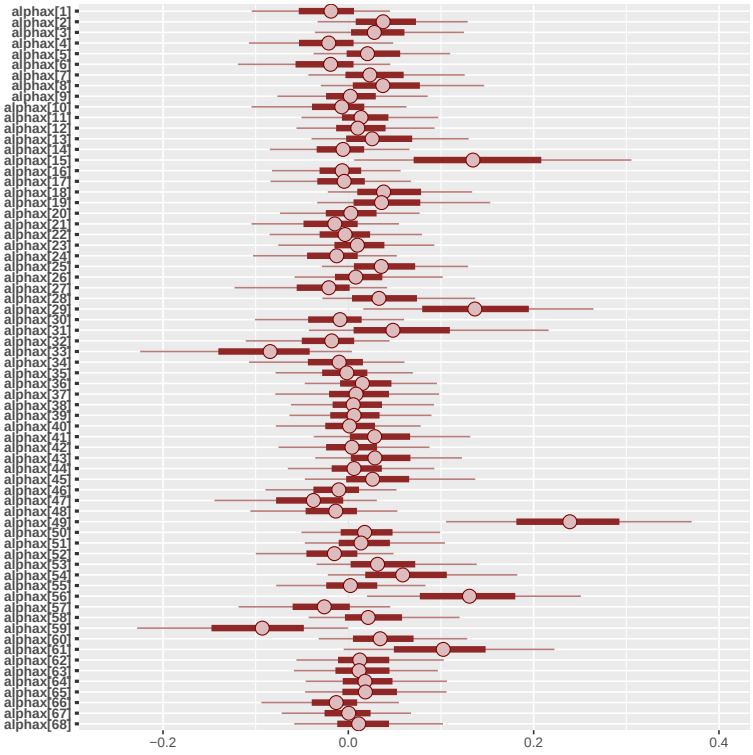

FIG. 3. Estimated posterior intervals for the instrument-exposure associations, α, based on our analysis with the final, complete, Bayesian model, that includes kinship information, family variable and parental protein levels.

(a continuous approximation of) the observed outcome variable distribution and the corresponding distribution generated from the posterior values of the unknown parameters of this final model. No signal of model misfit has been found (see Supplementary Material).

Gene IL12A (p35), together with gene IL12B (p40), encodes Interleukin 12 (abbreviated: IL12). IL12 is a pro-inflammatory cytokine, produced mainly by antigen presenting cells (abbreviated: APCs). It acts as an immunological playmaker by inducing Th1 cell differentiation from CD4+ naive T cells, interferon γ (abbreviated: IFN-γ) production and tumor necrosis factor-alpha (abbreviated: TNF-α) from T cells and natural killer (abbreviated: NK) cells (Aslani et al., 2017). A diagrammatic picture of the relevant pathway is shown in Figure 6. The hypothesised causal effect of IL12A on risk of MS might be mediated by the encoding of IL12 and the subsequent IL12-induced production of IFN-γ. In fact, IFN-γ is a major cytokine found in MS lesions, and it has been found that its levels are greatly increased during MS activity (Lees and Cross, 2007a). IL12-induced IFN-γ production is the key point in the Th1 immune responses induction and proliferation.

Furthermore, in murine models, IL12 has been shown to induce Substance P (SP) precursor mRNA in macrophages via STAT4 pathway (Arsenescu et al., 2005) and NK1R expression by both IL12 and IL18 stimulation via NFκB in T cells (Weinstock et al., 2003). SP has a demonstrated role in neuroimmune, autoimmune and inflammatory conditions, including MS (O'Connor et al., 2004; Kostyk et al., 1989). But while IL12 and IL23 are

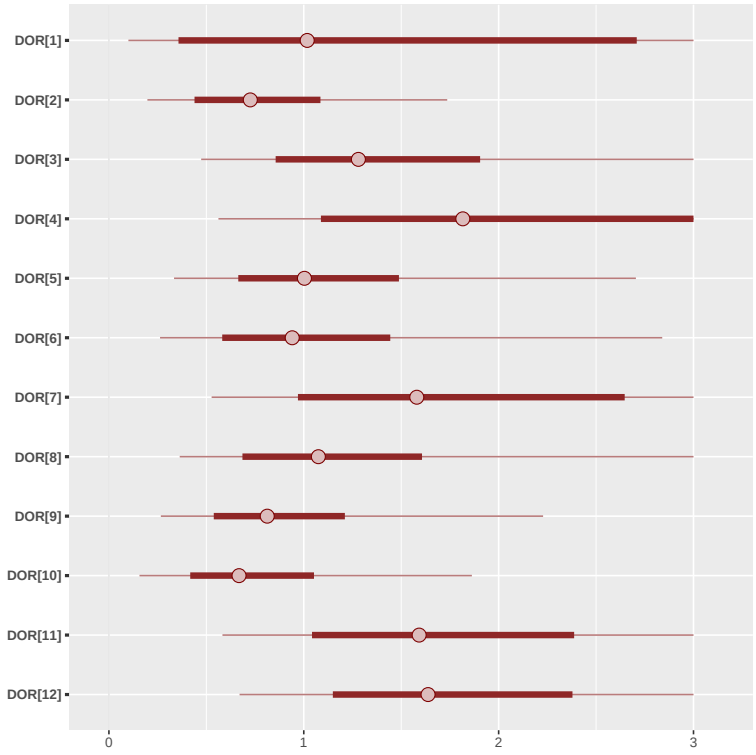

FIG. 4. Estimated direct causal effects of family membership on risk of MS, expressed on an odds-ratio scale, based on our analysis with the final, complete, Bayesian model, that includes kinship information, family indicator and parental protein levels.

pro-inflammatory cytokines, IL27 and IL35 are inhibitory cytokines. So, clearly, their immune balance is crucial for the modulation of immune function.

4. Discussion

We have extended the Bayesian MR framework of Berzuini and colleagues (Berzuini et al., 2018) for use in the analysis of pedigree data. MR has only rarely been applied to this class of data. Also, MR has been most frequently applied to the study of high-level exposures, such as as obesity (Conde et al., 2018; Mariosa et al., 2019), whereas our illustrative application deals with a molecular exposure. Some researchers appear confident that standard MR methods work equally well with molecular exposures, such as transcripts and proteins. Our early experiences in this area do not entirely corroborate this optimism, one reason being the intrinsic paucity of instruments at a molecular level. Although public bioinformatic repositories are sprawling with data, the number of available instruments for the analysis of causality at a molecular level is generally, and inevitably, poor due the the intrinsic nature of the studied mechanism. This makes MR analyses extremely vulnerable to the presence of confounding, not least because of possible, untestable, violations of the confounder independence assumption. MR analysis of pedigree data (as opposed to samples of unrelateds) promises robustness to confounding, and, for this reason, it presents itself as a useful tool for dealing with the information weakness we encounter in

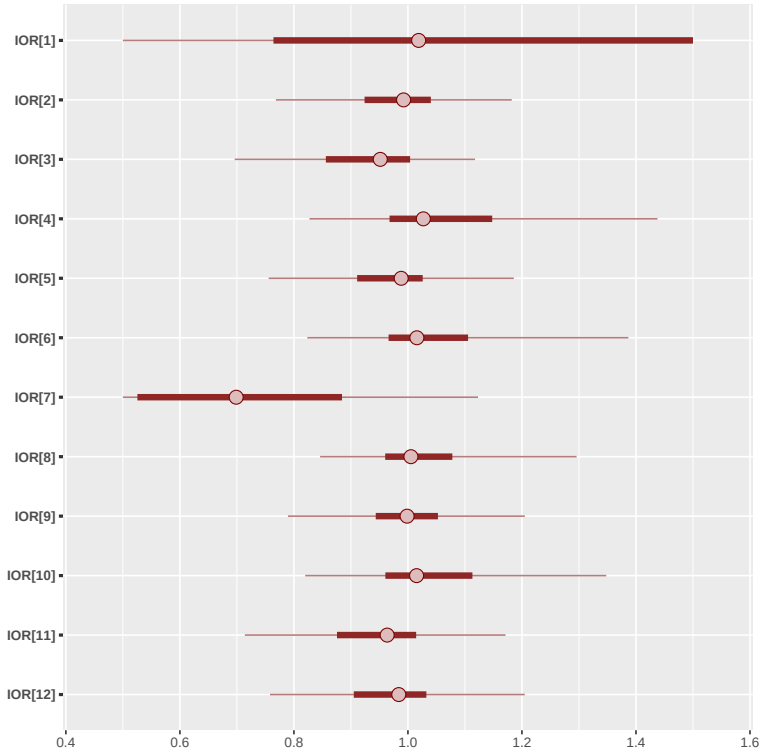

FIG. 5. Estimated indirect causal effects of family membership on risk of MS, expressed on an odds-ratio scale, based on our analysis with the final, complete, Bayesian model, that includes kinship information, family indicator and parental protein levels. We use the term "indirect" to signify the effect on MS risk exerted by membership to a particular family through the mediation of IL12A plasma level.

the study of causality at a molecular level. Motivated by these considerations, we have extended MR to work with pedigree data.

Results of our illustrative study point to the circulating level of protein IL12A as a potential cause of MS. While unexciting from a statistical significance viewpoint, our results match existing biological evidence. Interleukin 12 (IL12) is a pro-inflammatory cytokine, produced mainly by Antigen Presenting Cells (APCs). IL12 is a heterodimeric cytokine encoded by two separate genes, IL-12A (p35) and IL-12B (p40). It acts as an immunological playmaker inducing Th1 cell differentiation from CD4+ naive T cells, interferon γ (IFN-γ) production and tumor necrosis factor-alpha (TNF-α) from T cells and natural killer (NK) cells (Aslani et al., 2017; Katan, 1986). IFN-γ is a major cytokine found in MS lesions, and its levels are greatly increased during MS activity (Lees and Cross, 2007b). IFN-γ production induced by IL12 is the key point in the Th1 immune responses induction and proliferation. Furthermore, in murine models, IL-12 has been shown to induce precursor mRNA of Substance P (SP) in macrophages via STAT4 pathway (IL-12 induction of mRNA encoding substance P in murine macrophages from the spleen and sites of inflammation (Arsenescu et al., 2005)). In addition, both IL-12 and IL-18 stimulation induces NK1R expression via NFkB pathway in T cells (IL-18 and IL-12 signal through the NF-kappa B pathway to induce NK-1R expression on T cells (Weinstock et al., 2003)).

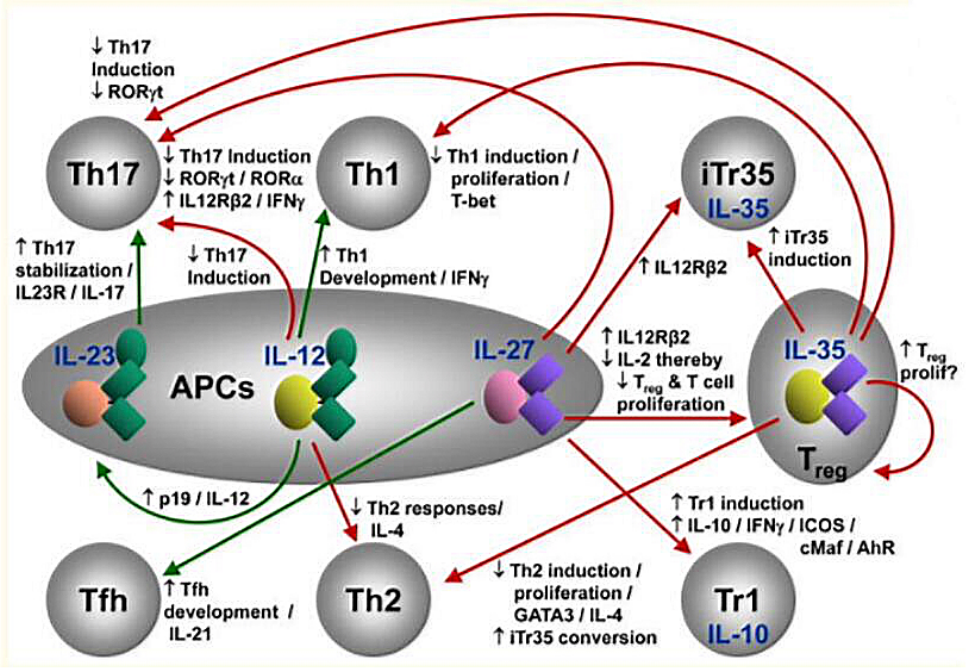

FIG. 6. IL12 family cytokines as a putative immunological link between IL12-A and MS.

SP has a demonstrated role in neuroimmune, autoimmune and inflammatory conditions, including MS (O'Connor et al., 2004; Kostyk et al., 1989). As shown in the figure 6, while IL-12 and IL-23 are pro-inflammatory cytokines, on the contrary IL-27 and IL-35 are inhibitory cytokines. So clearly, the immune balance of all the cytokines involved is crucial for the modulation of immune function where compensatory mechanisms can play a strategic role, that may explain the negative sign of the causal effect we found that is in contradiction with the expected increase of MS risk induced by IL12A.

From a statistical viewpoint, our IL12A data analysis illustrates a few important points. Firstly, because introduction of kinship information in the model accounts for the reduction in the number of "effective" individuals due to family correlation, it may result in an increased posterior uncertainty about the causal effect, with a reduction of evidence against the null causal hypothesis. Introduction of the family indicator may have a similar effect on the causal estimate, that of a greater posterior uncertainty, with a consequent further reduction of evidence of causality. Recall that family membership is a potential instrument-outcome confounder. The increase in posterior uncertainty consequent to introduction of the family indicator may thus be interpreted as an effect of the de-biasing. Our results suggest that our elaborations of the models tend to avoid over-optimistic results, which we believe to work in the direction of a healtier science. Parental protein information, introduced at the last model elaboration step, acted as instrumental, which resulted in an increase of evidence of causality.

MR has been traditionally applied to data from unrelated individuals. This is a pity, because MR analysis of family data is inherently more robust to population stratification

and heterogeneity than analysis of untelateds. We believe this property to help disentangle inheritable from environmental effects. A potentially fruitful idea is to collect data from unrelated individuals and then to collect further data from the parents of those individuals, for a joint analysis of the two data sources. Such a joint analysis can be performed via our proposed approach by treating parent-child triads as "families". Or one could use information from previous analyses of unrelateds in order to shape informative priors for an analysis of pedigree data along our proposed lines. Pedigree analysis might prove an invaluable tool for studying disease mechanism peculiarities of small, possibly native and isolated, populations. We are, in particular, thinking of small populations characterized by maverick disease patterns, that suffer from inadequate attention from the medical research community, perhaps outside the western "white" world.

Finally, on a more methodological note, we would emphasize the flexibility of a MCMC-powered Bayesian approach in MR, especially thanks to the possibility of straightforward elaboration of the basic MR model to accommodate extra relevant information and the straightforward handling of missing information.

We are at present working on an extension of the models discussed here to incorporate haplotype information.

REFERENCES

Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 8(3):559–575, 2007.

Razvan Arsenescu, Arthur Blum, Ahmed Metwali, David Elliott, and Joel Weinstock. Il-12 induction of mrna encoding substance p in murine macrophages from the spleen and sites of inflammation. *Journal of immunology (Baltimore, Md. : 1950)*, 174:3906–11, 05 2005. doi: 10.4049/jimmunol.174.7.3906.

S Aslani, N Jafari, and MR et al Javan. Epigenetic modifications and therapy in multiple sclerosis. *Neuromolecular Med.*, 19(1):11–23, 2017.

A.H. Beecham, N.A. Patsopoulos, D.K. Xifara, M.F. Davis, A. Kemppinen, C.J. Cotsapas, T.S. Shah, C.C.A. Spencer, D.R. Booth, A.N. Goris, A.B. Oturai, J. Saarela, B. Fontaine, B. Hemmer, C.F.C. Martin, F. Zipp, S. D'Alfonso, F. Martinelli-Boneschi, B.V. Taylor, H.F. Harbo, I. Kockum, J. Hillert, T.P. Olsson, M. Ban, J.R. Oksenberg, R.Q. Hintzen, L.F. Barcellos, C. Agliardi, L.S. Alfredsson, M. Alizadeh, C.A. Anderson, R. Andrews, H.B. Søndergaard, A.P. Baker, G. Band, S.E. Baranzini, N. Barizzone, J.C. Barrett, C. Bellenguez, L. Bergamaschi, L. Bernardinelli, A. Berthele, V. Biberacher, T.M.C. Binder, H. Blackburn, I.L. Bomfim, P. Brambilla, S.A. Broadley, B. Brochet, L. Brundin, D. Buck, H. Butzkueven, S.J. Caillier, W. Camu, W. Carpentier, P. Cavalla, E.G. Celius, I. Coman, G. Comi, L. Corrado, L. Cosemans, I. Cournu-Rebeix, B.A.C. Cree, D. Cusi, V. Damotte, G.L. Defer, S.R. Delgado, P. Deloukas, A. Di

Sapio, A.T. Dilthey, P.K. Donnelly, B.D. Dubois, M.E. Duddy, S.J. Edkins, I. Elovaara, F. Esposito, N. Evangelou, B. Fiddes, J.G. Field, A. Franke, C.O. Freeman, I.Y. Frohlich, D. Galimberti, C. Gieger, P.A.F.D. Gourraud, C. Graetz, A. Graham, V. Grummel, C. Guaschino, A. Hadjixenofontos, H.H. Hákonarson, C.A. Halfpenny, G. Hall, P. Hall, A. Hamsten, J. Harley, T.P. Harrower, C.P. Hawkins, G. Hellenthal, C.E.M. Hillier, J.C. Hobart, M. Hoshi, S.E. Hunt, M. Jagodić, I. Jelčić, A. Jochim, B. Kendall, Allan Kermode, T.J. Kilpatrick, K.J. Koivisto, I. Konidari, T. Korn, H.C. Kronsbein, C. Langford, M. Larsson, M.G. Lathrop, C. Lebrun-Frénay, J.S. Lechner-Scott, M. Lee, M.A. Leone, V. Leppä, G. Liberatore, B.A. Lie, C.M. Lill, M. Lindén, J. Link, F. Luessi, J.N. Lycke, F.M. Macciardi, S. Männistö, C.P. Manrique, R. Martin, V. Martinelli, D.M.L. Mason, G. Mazibrada, C. Mccabe, I.L. Mero, J. Mescheriakova, L. Moutsianas, K.M. Myhr, G. Nagels, R.S.J. Nicholas, P.C. Nilsson, F. Piehl, M. Pirinen, S.E. Price, H.L. Quach, M.I. Reunanen, W. Robberecht, N.P. Robertson, M.E. Rodegher, D.J. Rog, M. Salvetti, N.C. Schnetz-Boutaud, F.T. Sellebjerg, R.C. Selter, C.A. Schaefer, S. Shaunak, L. Shen, S. Shields, V. Siffrin, M. Slee, P.S. Sørensen, M. Sorosina, M. Sospedra, A. Spurkland, A. Strange, E. Sundqvist, V.N.S. Thijs, J.W. Thorpe, A.F. Ticca, P.J. Tienari, C.M.A. Van Duijn, E.M. Visser, S. Vucic, H. Westerlind, J.S. Wiley, A. Wilkins, J.F. Wilson, J. Winkelmann, J.P. Zajicek, E. Zindler, J.L. Haines, M.A. Pericak-Vance, A.J. Ivinson, G.J. Stewart, D.D. Hafler, S.L. Hauser, A.D. Compston, G.A.T. Mcvean, P.L. De Jager, S.J. Sawcer, and J.L. Mccauley. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, 45(11):1353–1362, 2013. ISSN 1061-4036. doi: 10.1038/ng.2770.

Carlo Berzuini, Hui Guo, Stephen Burgess, and Luisa Bernardinelli. A Bayesian approach to Mendelian randomization with multiple pleiotropic variants. *Biostatistics*, 08 2018. ISSN 1465-4644. doi: 10.1093/biostatistics/kxy027. URL `https://doi.org/10.1093/biostatistics/kxy027`.

Ben Brumpton, Eleanor Sanderson, Fernando Pires Hartwig, Sean Harrison, Gunnhild Åberge Vie, Yoonsu Cho, Laura D Howe, Amanda Hughes, Dorret I Boomsma, Alexandra Havdahl, John Hopper, Michael Neale, Michel G Nivard, Nancy L Pedersen, Chandra A Reynolds, Elliot M Tucker-Drob, Andrew Grotzinger, Laurence Howe, Tim Morris, Shuai Li, , Wei-Min Chen, Johan Håkon Bjørngaard, Kristian Hveem, Cristen Willer, David M Evans, Jaakko Kaprio, George Davey Smith, Bjørn Olav Åsvold, Gibran Hemani, and Neil M Davies. Within-family studies for mendelian randomization: avoiding dynastic, assortative mating, and population stratification biases. *bioRxiv*, 2019. doi: 10.1101/602516. URL `https://www.biorxiv.org/content/early/2019/04/09/602516`.

B Carpenter, A Gelman, and et al. Hoffman, MD. `Stan`: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32, 2017.

C.M. Carvalho, N.G. Polson, and J.G Scott. The horseshoe estimator for sparse signals. *Biometrika*, page asq017, 2010.

Susana Conde, Xiaoguang Xu, Hui Guo, Markus Perola, Teresa Fazia, Luisa Bernardinelli, and Carlo Berzuini. Mendelian randomisation analysis of clustered causal effects of body

mass on cardiometabolic biomarkers. *BMC Bioinformatics*, 19-S(7):13–22, 2018. doi: 10.1186/s12859-018-2178-2. URL `https://doi.org/10.1186/s12859-018-2178-2`.

Cris S Constantinescu, Maria Wysocka, Brendan Hilliard, Elvira S Ventura, Ehud Lavi, Giorgio Trinchieri, and Abdolmohamad Rostami. Antibodies against il-12 prevent superantigen-induced and spontaneous relapses of experimental autoimmune encephalomyelitis. *The Journal of Immunology*, 161(9):5097–5104, 1998.

Panayiota Constantinou and A. Philip Dawid. Extended conditional independence and applications in causal inference. *Ann. Statist.*, 45(6):2618–2653, 12 2017. doi: 10.1214/16-AOS1537. URL `https://doi.org/10.1214/16-AOS1537`.

A.P. Dawid. Conditional independence in statistical theory (with Discussion). *J. Roy. Statist. Soc. B*, 41:1–31, 1979.

A.P. Dawid. Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association*, 95:407–448, 2000.

A.P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002. Corrigenda, *ibid.*, 437.

T Fazia, R Pastorino, L Foco, L Han, M Abney, A Beecham, A Hadjixenofontos, Hui Guo, D Gentilini, C Papachristou, P.P Bitti, A Ticca, Carlo Berzuini, J McCauley, and L Bernardinelli. Investigating multiple sclerosis genetic susceptibility on the founder population of east-central sardinia via association and linkage analysis of immune-related loci. *Multiple Sclerosis Journal*, 2017. ISSN 1352-4585. doi: 10.1177/1352458517732841.

J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society, Series A*, 182:389–402, 2019. doi: 10.1111/rssa.12378.

D. Geiger, T. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.

Yang J, Manolio TA, and etal Pasquale LR. Genome partitioning of genetic variation for complex traits using common snps. *Nature Genetics Nature Research??????*, 43: 519–525, 2011. URL `http://www.nature.com/doifinder/10.1038/ng.823`.

Hamidreza Jahanbani-Ardakani, Fereshteh Alsahebfosoul, Saeed-Reza Moshfeghi, Behzad Mahaki, Masoud Etemadifar, Seyed-Hossein Abtahi, and Mahboobeh Fereidan-Esfahani. Serum level of interleukin 12 in patients with multiple sclerosis. *International Journal of Neuroscience*, 129(2):207–208, 2019.

M.B. Katan. Apolipoprotein e isoforms, serum cholesterol and cancer. *The Lancet*, 327 (8479):507–508, 3 1986. ISSN 0140-6736. doi: 10.1016/S0140-6736(86)92972-7.

S.K. Kostyk, N.W. Kowall, and S.L. Hauser. Substance p immunoreactive astrocytes are present in multiple sclerosis plaques. *Brain Research*, 504(2):284 – 288, 1989. ISSN 0006-8993. doi: https://doi.org/10.1016/0006-8993(89)91369-3. URL `http://www.sciencedirect.com/science/article/pii/0006899389913693`.

Steffen L. Lauritzen. *Graphical models*. Clarendon Press, 1996. ISBN 0198522193.

Jason R. Lees and Anne Cross. A little stress is good: Ifn-γ, demyelination, and multiple sclerosis. *Journal of Clinical Investigation*, 117(2):297–299, 2 2007a. ISSN 0021-9738. doi: 10.1172/JCI31254.

Jason R Lees and Anne H Cross. A little stress is good: Ifn-γ, demyelination, and multiple sclerosis. *The Journal of clinical investigation*, 117(2):297–299, 2007b.

D Mariosa, R Carreras-Torres, RM Martin, M Johansson, and P Brennan. What can mendelian randomization tell us about causes of cancer? *Int J Epidemiol*, 48(3):816–821, 2019.

Peter Nilsson, Linda Paavilainen, Karin Larsson, Jenny Ödling, Mårten Sundberg, Ann-Catrin Andersson, Caroline Kampf, Anja Persson, Cristina Al-Khalili Szigyarto, Jenny Ottosson, Erik Björling, Sophia Hober, Henrik Wernérus, Kenneth Wester, Fredrik Pontén, and Mathias Uhlen. Towards a human proteome atlas: High-throughput generation of mono-specific antibodies for tissue profiling. *PROTEOMICS*, 5(17):4327–4337, 2005. doi: 10.1002/pmic.200500072. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.200500072`.

T.M. O'Connor, J. O'Connell, D.I. O'Brien, T. Goode, C.P. Bredin, and F. Shanahan. The Role of Substance P in Inflammatory Disease., journal=Journal of Cellular Physiology, volume=201, pages=167–180. 2004.

Juho Piironen and Aki Vehtari. *On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior*, volume 54 of *Proceedings of Machine Learning Research*. PMLR, April 2017. URL `http://proceedings.mlr.press/v54/piironen17a.html`.

José C. Pinheiro and Douglas M. Bates. *Mixed-effects models in S and S-PLUS*. New York: Springer, 2000.

Charles M. Poser, Donald W. Paty, Labe Scheinberg, W. Ian McDonald, Floyd A. Davis, George C. Ebers, Kenneth P. Johnson, William A. Sibley, Donald H. Silberberg, and Wallace W. Tourtellotte. New diagnostic criteria for multiple sclerosis: Guidelines for research protocols. *Annals of Neurology*, 13(3):227–231, 1983. doi: 10.1002/ana.410130302. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.410130302`.

Michael Rentzos, Chryssoula Nikolaou, Antonis Rombos, M Eleftheria Evangelopoulos, Evangelia Kararizou, George Koutsis, Margarita Zoga, Antonis Dimitrakopoulos, Anthousa Tsoutsou, and Costas Sfangos. Effect of treatment with methylprednisolone on the serum levels of il-12, il-10 and ccl2 chemokine in patients with multiple sclerosis in relapse. *Clinical neurology and neurosurgery*, 110(10):992–996, 2008.

Jochen M. Schwenk, Johan Lindberg, Mårten Sundberg, Mathias Uhlén, and Peter Nilsson. Determination of binding specificities in highly multiplexed bead-based assays for antibody proteomics. *Molecular & Cellular Proteomics*, 6(1):125–132, 2007. ISSN 1535-9476. doi: 10.1074/mcp.T600035-MCP200. URL `https://www.mcponline.org/content/6/1/125`.

Jochen M. Schwenk, Marcus Gry, Rebecca Rimini, Mathias Uhlén, and Peter Nilsson. Antibody suspension bead arrays within serum proteomics. *Journal of Proteome Research*, 7(8):3168–3179, 2008. doi: 10.1021/pr700890b. URL `https://doi.org/10.1021/pr700890b`. PMID: 18588325.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search, 2nd Edition*, volume 1 of *MIT Press Books*. The MIT Press, March 2001. ISBN ARRAY(0x439ee6b0). URL `https://ideas.repec.org/b/mtp/titles/0262194406.html`.

B. Sun, J. C. Maranville, J. Peters, D. Stacey, J. Staley, J. Blackshaw, S. Burgess, and et al. Genomic atlas of the human plasma proteome. *Nature*, 558:73–79, 2018. doi: https://doi.org/10.1038/s41586-018-0175-2.

Lin Sun, Chang He, Lekha Nair, Justine Yeung, and Charles E Egwuagu. Interleukin 12 (il-12) family cytokines: Role in immune pathogenesis and treatment of cns autoimmune disease. *Cytokine*, 75(2):249–255, 2015.

Joel V. Weinstock, Arthur Blum, Ahmed Metwali, David Elliott, and Razvan Arsenescu. Il-18 and il-12 signal through the nf-κb pathway to induce nk-1r expression on t cells. *The Journal of Immunology*, 170(10):5003–5007, 2003. ISSN 0022-1767. doi: 10.4049/jimmunol.170.10.5003. URL `https://www.jimmunol.org/content/170/10/5003`.

Wang X, Tucker NR, and et al Rizki G. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *Elife*, 5, 2016.

OO Yavorska and Stephen Burgess. `MendelianRandomization`: an `r` package for performing mendelian randomization analyses using summarized data. *International Journal of Epidemiology*, 46:1734–1739, 2017.

APPENDIX A

Stan code

```
BayesianMR <-'
 data {
 int<lower=0> nobs;              number of individuals with non-missing value of X
 int<lower=0> nmis;              number of individuals with missing value of X
 int<lower=0> nfam;                                        total number of families
 int<lower=0> N;              total number of sample individuals (=NOBS+NMIS)
 int<lower=0> J;                                        total number of instruments
 matrix[N,J] Z;      matrix of standardized (mean= 0, SD= 1) values of instruments
 matrix[N,N] pedigree_matrix;                  kinship matrix over whole sample
 vector[nobs] Xobs;      observed X values, transformed to a 0-mean-1-SD variables
 int Y[N];                                          vector of (0,1) disease indicators
 vector[N] PROTEINMADRE;                          measured protein level in mother
 vector[N] PROTEINPADRE;                          measured protein level in father
 real<lower=0> betasimsd;
 real<lower=0> cauchysd;      scale parameter for Cauchy prior on causal parameter
 real<lower=1> nu_global;             degrees of freedom for the half-t prior for tau
 real<lower=1> nu_local;          df half-t priors for the LAMBDAs (1→horseshoe)
 matrix[N,nfam] FAM;      matrix of standardized (mean= 0,SD= 1) family indicators
 vector[N] muY;                                                        auxiliary
 }

transformed data{
 matrix[N,N] L_pedigree_matrix = cholesky_decompose(pedigree_matrix);
 }

parameters {
 real<lower=0,upper=1> frazionepleio;          expected fraction pleiotrop. instr.
 real <lower=0> sigmax;          lik-unidentifiable SD of measurement error on X
 real <lower=0> sigmaalpha;                              SD of ALPHAX hyperprior
 real theta;                                  causal effect of inferential interest
 real <lower=-1,upper=1> alphaMADRE;
 real <lower=-1,upper=1> alphaPADRE;
 real omegay;                                          intercept of the model for Y
 real deltax;               effect of U on X, and covariance betw X- and Y-errors
 vector[N] u;                                  individual-specific confounder values
 vector[J] alphax;                               effects of instruments on exposure
 vector[nmis] Xmis;                              unobserved subset of values of X
 vector[nfam] gammafamx;
 vector[nfam] gammafamy;
 vector[N] correction;
```

Auxiliary variables that define the global and local parameters:

```
vector[J] z;
real<lower=0> r1_global;
real<lower=0> r2_global;
vector<lower=0>[J] r1_local;
vector<lower=0>[J] r2_local;
}


  transformed parameters {
```

Half-t priors for the lambdas:

```
vector[J] beta;                          unknown pleiotropic effects in real dataset
real<lower=0> tau;                                    global shrinkage parameter
vector<lower=0>[J] lambda;                            local shrinkage parameter
real<lower=0> m0;
real<lower=0> scale_global;
vector[N] Xcompleto;
lambda = r1_local .* sqrt(r2_local);
tau= r1_global * sqrt(r2_global);
beta = z .* lambda*tau;
m0=floor(J*frazionepleio);           expected number of large pleiotropic effects
scale_global= 2*m0/(sqrt(N)*(J-m0));
Xcompleto[1:nobs]= Xobs;
Xcompleto[(nobs+1):N] = Xmis;
}


  model {
frazionepleio ~ uniform(0.1,0.9);
z~ normal(0,1);
r1_local~ normal(0.0,1.0);
r2_local~ inv_gamma(0.5*nu_local,0.5*nu_local);
r1_global~ normal(0.0,scale_global);          required for half-t prior for TAU
r2_global~ inv_gamma(0.5*nu_global,0.5*nu_global);                  as above
```

Model for observed values of X

```
Xobs     ~ normal(
     FAM[1:nobs,]*gammafamx                                  family → X
     +Z[1:nobs,]*alphax                                  instruments → X
     +PROTEINMADRE[1:nobs]*alphaMADRE              maternal protein level → X
     +PROTEINPADRE[1:nobs]*alphaPADRE
```

```
     +u[1:nobs]*deltax,                          unknown confounder → X
sigmax);


                                                 Model for unobserved values of X,
Xmis~ normal(                                    to be imputed as part of inference
     FAM[(nobs+1):N,]*gammafamx+Z[(nobs+1):N,]*alphax
     +PROTEINMADRE[(nobs+1):N]*alphaMADRE
     +PROTEINPADRE[(nobs+1):N]*alphaPADRE
     +u[(nobs+1):N]*deltax,
sigmax);


                                                          Observation model for Y
correction ~ multi_normal_cholesky(muY, L_pedigree_matrix);
Y ~ bernoulli_logit(omegay+Z*beta +FAM*gammafamy
     +Xcompleto*theta +u +correction);


                                                                          Prior
theta ~ cauchy(0,cauchysd);
for(n in 1:N){
u[n] ~ normal(0,1);}
alphaMADRE ~ uniform(-1,1);
alphaPADRE ~ uniform(-1,1);
for(h in 1:nfam){
  gammafamx[h] ~ cauchy(0,cauchysd);
  gammafamy[h] ~ cauchy(0,cauchysd);}
  for(k in 1:J){
    alphax[k]~ double_exponential(0, sigmaalpha);          note 0 mean
}
}
'
```

TABLE 5. Additional results from our final model. For each of the 12 families represented in our data, this table reports the estimated effect that being a member of that family has on MS risk, by distinguishing between the direct and the indirect (=mediated by changes in the level of circulating IL12A protein) components of the effect. See rigorous definition of direct and indirect effect in the Methods section.

FAMILY	PERCENTILES OF POSTERIOR DISTRIBUTION OF EFFECT				
	5	25	50	75	95
FAMILY-SPECIFIC **Indirect** CAUSAL EFFECT ON RISK OF MS (ODDS RATIO)					
family 2	0.77	0.92	0.99	1.04	1.18
family 3	0.70	0.86	0.95	1.00	1.12
family 4	0.83	0.97	1.03	1.15	1.44
family 5	0.76	0.91	0.99	1.03	1.19
family 6	0.82	0.97	1.02	1.11	1.39
family 7	0.31	0.53	0.70	0.88	1.12
family 8	0.85	0.96	1.01	1.08	1.30
family 9	0.79	0.94	1.00	1.05	1.21
family 10	0.82	0.96	1.02	1.11	1.35
family 11	0.71	0.88	0.96	1.01	1.17
family 12	0.76	0.91	0.98	1.03	1.21
FAMILY-SPECIFIC **Direct** CAUSAL EFFECT ON RISK OF MS (ODDS RATIO)					
family 2	0.20	0.44	0.73	1.09	1.74
family 3	0.47	0.86	1.28	1.90	3.33
family 4	0.56	1.09	1.82	3.43	10.04
family 5	0.33	0.66	1.00	1.49	2.70
family 6	0.26	0.58	0.94	1.44	2.84
family 7	0.53	0.97	1.58	2.65	5.67
family 8	0.36	0.69	1.07	1.61	3.09
family 9	0.27	0.54	0.81	1.21	2.23
family 10	0.16	0.42	0.67	1.05	1.86
family 11	0.58	1.04	1.59	2.39	4.28
family 12	0.67	1.15	1.64	2.38	4.24