Multiple Sclerosis and Related Disorders 24 (2018) 135–141

Contents lists available at ScienceDirect

Multiple Sclerosis and Related Disorders

journal homepage: www.elsevier.com/locate/msard

Incorporating machine learning approaches to assess putative environmental risk factors for multiple sclerosis

Ellen M. Mowry[a,1,*], Anna K. Hedström[b,c,1], Milena A. Gianfrancesco[d], Xiaorong Shao[d], Catherine A. Schaefer[e], Ling Shen[e], Kalliope H. Bellesis[e], Farren B.S. Briggs[f], Tomas Olsson[g], Lars Alfredsson[b,2], Lisa F. Barcellos[d,2]

[a] Johns Hopkins University, 600N. Wolfe Street, Pathology 627, Baltimore 21287, MD, USA
[b] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[c] Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
[d] University of California, Berkeley, CA, USA
[e] Kaiser Permanente Division of Research, Oakland, CA, USA
[f] Case Western Reserve University, Cleveland, OH, USA
[g] Karolinska Institutet at Karolinska University Hospital, Solna, Sweden

ARTICLE INFO

Keywords:

Multiple sclerosis
Epidemiology
Gene-environment interaction
Occupational exposure
Environmental exposure

ABSTRACT

Background: Multiple sclerosis (MS) incidence has increased recently, particularly in women, suggesting a possible role of one or more environmental exposures in MS risk. The study objective was to determine if animal, dietary, recreational, or occupational exposures are associated with MS risk.

Methods: Least absolute shrinkage and selection operator (LASSO) regression was used to identify a subset of exposures with potential relevance to disease in a large population-based (Kaiser Permanente Northern California [KPNC]) case-control study. Variables with non-zero coefficients were analyzed in matched conditional logistic regression analyses, adjusted for established environmental risk factors and socioeconomic status (if relevant in univariate screening), ± genetic risk factors, in the KPNC cohort and, for purposes of replication, separately in the Swedish Epidemiological Investigation of MS cohort. These variables were also assessed in models stratified by *HLA-DRB1*15:01* status since interactions between risk factors and that haplotype have been described.

Results: There was a suggestive association of pesticide exposure with having MS among men, but only in those who were positive for *HLA-DRB1*15:01* (OR pooled = 3.11, 95% CI 0.87, 11.16, p = 0.08).

Conclusions: While this finding requires confirmation, it is interesting given the association between pesticide exposure and other neurological diseases. The study also demonstrates the application of LASSO to identify environmental exposures with reduced multiple statistical testing penalty. Machine learning approaches may be useful for future investigations of concomitant MS risk or prognostic factors.

1. Introduction

Multiple sclerosis (MS) is a complex neurologic disorder for which susceptibility is conferred by both genetic and environmental factors. Variation within the human leukocyte antigen (*HLA*)-*DRB1*15:01* allele exerts the greatest genetic effect on MS risk, and many other loci with very modest effects have been identified (Hafler et al., 2007; Patsopoulos et al., 2011; Sawcer et al., 2011). However, the substantial MS discordance in monozygotic twins suggests that factors in the environment influence disease risk (Hawkes and Macgregor, 2009). Lower vitamin D levels, obesity, exposure to tobacco cigarette smoke and Epstein-Barr virus have shown consistent associations with the development of MS (Ascherio, 2013).

The incidence of MS, particularly in women, has increased in the past several decades (Orton et al., 2006; Trojano et al., 2012), implicating new or changing environmental MS risk factors. While lower vitamin D levels have been proposed to explain some of this shift (Saltzer et al., 2012), recent changes in sun exposure behaviors may not

* Corresponding author.
E-mail address: Emowry1@jhmi.edu (E.M. Mowry).
[1] Drs. Mowry and Hedström are co-first authors.
[2] Dr. Alfredsson and Barcellos are co-senior authors.

https://doi.org/10.1016/j.msard.2018.06.009
Received 16 February 2018; Received in revised form 7 May 2018; Accepted 15 June 2018

have been substantial enough to dramatically change vitamin D status (Robinson et al., 1997). Potential or known toxic agents, some of which have been associated with other neurodegenerative disorders (e.g. pesticides and Parkinson disease [PD]) (Wang et al., 2011), have been inadequately explored for their association with MS risk. Further, whether other occupational, recreational, animal or dietary exposures influence the risk of the disease is unclear.

Determining key predictors for analyses when detailed genetic and environmental exposure data are available is challenging due to concerns about multiple testing and correlations between predictors. Machine learning methods present a wide array of scalable methods to analyze complex, high-dimensional datasets comprised of a significant number of variables with various degrees of dependencies (Bellinger et al., 2017). These techniques have been utilized to address the conundrum of selecting putative predictors for evaluation of their role in disease status in other patient populations (Jamal et al., 2016), allowing for improved modeling in several cases. For example, to evaluate the relationship of dietary nutrients with coronary artery calcification, one study conducted least absolute shrinkage and selection operator regression (LASSO) in order to take into account the strong co-linearity of nutrients and was thus able to identify three micronutrients of interest (Machado et al., 2018). Several recent studies have demonstrated that machine learning techniques lead to more accurate predictive models. Among patients with early, pedunculated colorectal cancer, LASSO out-performed two traditional statistical models in predicting which patients have an increased risk of metastases and thus required certain types of subsequent care (Backes et al., 2018). In assessing risk of inpatient mortality, LASSO improved the prediction model compared to traditional stepwise logistic regression (Schwartz et al., 2018). A final study of note evaluated predictors of cardiac death, taking into account a number of clinical and paraclinical data, and found that machine learning models, including LASSO, outperformed traditional logistic regression in establishing the area under the receiver operating characteristic curve (Haro Alonso et al., 2018). In a case-control study in the Kaiser Permanente Northern California (KPNC) health plan, the LASSO method was used to select environmental exposures for further evaluation in multivariate models in both the KPNC and the Swedish Epidemiological Investigation in Multiple Sclerosis (EIMS) case-control studies. Given that MS incidence is greater in women and because sex itself may determine the degree of exposure to several factors assessed herein (e.g. nail polish, hair dye), the analyses were performed separately for women and men. Further, LASSO-identified putative risks were evaluated in models stratified by *HLA-DRB1*15:01* status since interactions between it and other environmental risk factors have been identified (Hedström et al., 2014a, 2014b; Sundqvist et al., 2012).

2. Materials and methods

2.1. Study population

2.1.1. KPNC

The study protocol was approved by the Institutional Review Boards (IRBs) of KPNC and the University of California, Berkeley. The study population included white, non-Hispanic people with MS and matched controls, identified through electronic medical records, who were members of KPNC, which comprises approximately 25–30% of the population in the 22-county region in northern California it serves. KPNC members are demographically similar to the general population except that extremes of income are underrepresented (Krieger, 1992).

To be eligible as a case, a subject had at least one outpatient MS diagnosis by a neurologist and had to be white, non-Hispanic, currently aged 18–69 years, and a member of KPNC at the time of contact. Diagnoses were validated by chart review and radiology (MRI) and pharmacy records, according to McDonald criteria (McDonald et al., 2001; Polman et al., 2005). In addition, the treating neurologist of each potential case was contacted to gain approval to contact the person; those who did not have MS, were deemed too ill or impaired to participate, or who were no longer KPNC members were thus excluded. The study began in mid-2006. The data freeze for the current study occurred in August 2014, by which time a total of 3293 potential MS cases had been reviewed by KPNC neurologists, who approved contact with 2823 (86%).

Potential cases were initially contacted by mail, followed by a telephone call from project staff, who explained the study procedures and confirmed eligibility; those individuals made an appointment to complete a computer-assisted telephone interview (CATI; see below). Participants were mailed a consent form and a blood sample collection kit after the interview was complete; they returned the signed consent form by mail and took the blood collection kit to a KPNC laboratory, where the blood was drawn and the tubes were shipped with next-day delivery for processing. If a blood sample could not be obtained, participants completed and returned a saliva collection kit.

Controls were KPNC members without a diagnosis of MS or related conditions (transverse myelitis, optic neuritis, or demyelination disease; ICD9 codes: 340, 341.0, 341.1, 341.2, 341.20, 341.21, 341.22, 341.8, 341.9, 377.3, 377.30, 377.39, and 328.82) randomly selected from current members using electronic medical records. Controls were matched to individual cases on sex, birth date (± 1 year), race/ethnicity (self-identified), and zip code of the residence of the case. Controls were screened by telephone to confirm eligibility; another was selected if the first control was ineligible or refused. Controls were initially contacted by mail, followed by a phone call, to explain the study and procedures. Controls who agreed to participate completed the CATI, the mailed consent form, and the saliva collection kit. At the time of the data freeze, the average participation rate was approximately 80% for cases and 66.4% for controls. The final dataset included 1,158 cases and 1,179 matched controls.

The CATI included questions designed to capture exposure (prior to MS onset) regarding dietary, recreational, occupational/habitation, and pet exposures, as well as known environmental risk factors (Ascherio, 2013). A history of infectious mononucleosis was used as a proxy for prior Epstein-Barr virus infection since viral serostatus prior to MS onset was not available. A history of taking vitamin D was considered a proxy for vitamin D status, while a history of smoking cigarettes (at least one cigarette per day for at least one month) was used as a measure of cigarette exposure. Body size at age 10 was reported as a proxy of body mass index, as previously reported (Gianfrancesco et al., 2014). Cases answered questions about their MS, including the timing of onset and of diagnosis. The onset year was determined as the year of the first symptom reported by the cases; this was considered the reference year for matched controls. The CATI tracked and supplied the appropriate reference year for each case and matched control.

For DNA extraction, Oragene kits were used to collect saliva; kits were processed according to the manufacturer's protocol. DNA was extracted after whole blood collection and processed using standard methods. Genotyping was performed as previously described (Barcellos et al., 20062009). Genetic data were used to calculate a weighted genetic risk score (wGRS) as described (DeJager et al., 2009).

2.1.2. EIMS

The EIMS study, approved by the Regional Ethical Review Board at Karolinska Institutet, has as its study base the Swedish population aged 16–70 years and has led to multiple publications regarding environmental factors influencing MS risk (Hedström et al., 2014a, 2014b; Sundqvist et al., 2012; Hedström et al., 2016a, 2016b2015, 2014). Incident MS cases were recruited at 40 study centers, including all university hospitals in Sweden. All MS cases were diagnosed by a neurologist using specified criteria (Thompson et al., 2000). Two controls were randomly selected per case from the national population register, matched by age (in five-year strata), sex and residential area.