Received: 29 May 2020 | Revised: 13 November 2020 | Accepted: 19 November 2020

DOI: 10.1002/sim.8843

RESEARCH ARTICLE

Statistics in Medicine WILEY

Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship

Alice J. Sommer[1,2,3] | **Emmanuelle Leray**[4] | **Young Lee**[1] | **Marie-Abèle C. Bind**[1]

[1]Department of Statistics, Harvard University, Cambridge, Massachusetts

[2]Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine Ludwig-Maximilians-University München, Munich, Germany

[3]Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany

[4]University of Rennes, EHESP French School of Public Health, REPERES Pharmacoepidemiology and Health Services Research EA, 7449, Rennes, France

Correspondence
Marie-Abèle C. Bind, Department of Statistics, Harvard University, One Oxford Street, Cambridge, MA 02138.
Email: ma.bind@mail.harvard.edu

Funding information
John Harvard Distinguished Science Fellows Program; NIH Office of the Director, Grant/Award Number: DP5OD021412

When addressing environmental health-related questions, most often, only observational data are collected for ethical or practical reasons. However, the lack of randomized exposure often prevents the comparison of similar groups of exposed and unexposed units. This design barrier leads the environmental epidemiology field to mainly estimate associations between environmental exposures and health outcomes. A recently developed causal inference pipeline was developed to guide researchers interested in estimating the effects of plausible hypothetical interventions for policy recommendations. This article illustrates how this multistaged pipeline can help environmental epidemiologists reconstruct and analyze hypothetical randomized experiments by investigating whether an air pollution reduction intervention decreases the risk of multiple sclerosis relapses in Alsace region, France. The epidemiology literature reports conflicted findings on the relationship between air pollution and multiple sclerosis. Some studies found significant associations, whereas others did not. Two case-crossover studies reported significant associations between the risk of multiple sclerosis relapses and the exposure to air pollutants in the Alsace region. We use the same study population as these epidemiological studies to illustrate how appealing this causal inference approach is to estimate the effects of hypothetical, but plausible, environmental interventions.

KEYWORDS
causal inference, environmental epidemiology, matching, multiple sclerosis, observational data

1 | INTRODUCTION

The major reason for the confidence in randomized experiments is the objectivity of the decisions for exposure assignment to compare treated and control units with similar pre-exposure covariates. Following the logic of the Rubin Causal Model, the appealing features of randomized experiments can be transposed to observational studies to provide transparent cause and effect interpretations of statistical analyses.[1,2] These ideas should be particularly appealing to environmental epidemiology, a field for which randomized experiments are most often unethical or impractical. Bind and Rubin[3] present, with a simple illustration, a multistaged causal inference pipeline aiming at revealing results that could have been obtained

TABLE 1 Multiple sclerosis patient characteristics

	Total		Cluster 1		Cluster 2	
	n = 353		n = 146		n = 207	
Sex	M	F	M	F	M	F
n	98 (28%)	255 (72%)	38 (26%)	108 (74%)	60 (29%)	147 (71%)
MS form (at last info.)						
Relapsing-remitting (n)	80 (23%)	216 (61%)	27 (18%)	73 (50%)	53 (26%)	143 (69%)
Secondary progressive (n)	18 (5%)	39 (11%)	11 (8%)	35 (24%)	7 (3%)	4 (2%)
Relapses per patient (mean)	1	2	2	2	1	1
SD	(2)	(2)	(2)	(2)	(2)	(2)
Onset—inclusion gap (mean in years)	3	4	7	9	0.1	0.1
SD	(6)	(7)	(8)	(8)	(0.5)	(0.5)
Age at MS clinical onset (mean in years)	31	31	33	33	29	30
SD	(11)	(10)	(9)	(12)	(10)	(11)
Age at study inclusion (mean in years)	36	36	41	42	30	31
SD	(11)	(12)	(8)	(11)	(9)	(10)
Follow-up since inclusion (mean in years)	6	6	9	9	4	4
SD	(4)	(4)	(1)	(1)	(3)	(3)
Type 1—*prevalent cases* (n)	35 (10%)	97 (27%)	31 (21%)	88 (60%)	4 (2%)	9 (4%)
Type 2—*incident cases* (n)	63 (18%)	158 (45%)	7 (5%)	20 (14%)	56 (27%)	138 (67%)

by an experiment with a plausible, randomly assigned, environmental intervention. A recent study partially followed this pipeline: the authors constructed the nonrandomized data such that they mimic random weather variations and estimated the effects of weather variations on violent crimes.[4] However, the study examined weather variations that cannot be interpreted as plausible interventions, thereby omitting the first, conceptual, stage of the pipeline, essential for providing recommendations.

A few researchers have emphasized the importance of focusing on well-defined interventions in terms of potential outcomes and on relying on the Rubin Causal Model to assess causality in air pollution epidemiology.[3,5-9] For years, Rubin has argued for a design stage[10] and for a conceptual stage,[2,11] which was formulated more explicitly recently by Bind and Rubin. Here, our objective is to provide epidemiologists with a practical and thorough application of the causal pipeline proposed by Bind and Rubin and simultaneously assess an important causal question with a complex data structure.

There is an increasing number of epidemiological studies focusing on the link between air pollution and neurological outcomes, including multiple sclerosis (MS) relapses.[12,13] MS is a demyelinating disease damaging nerve cells, giving rise to the inability of the nervous system to communicate. MS patients occasionally experience relapses. Relapses are characterized as episodes of neurological symptoms (eg, loss of vision, pain in body parts) that occur for at least 24 hours and happen at least 30 days after any previous episode began.[14] The causes of MS disease onset and the risk factors of relapses occurrence are unclear but many research efforts are focusing on the influence of environmental factors on MS.[15,16] Several studies reported associations between air pollutants and MS[17-23] and two studies failed to reject the null hypothesis,[24,25] (see Web Table 1). However, study design and methodological limitations prevent a causal interpretation for these associations.

Our illustration uses a study population already studied by Roux et al[21] and Jeanjean et al,[23] who observed positive associations between air pollutants and MS relapses risk. Both concluded that further research presenting

causal relationships is needed before taking preventive environmental actions for MS patients. These studies rely on a case-crossover strategy[26] that examines whether the patient was exposed to some unusual air pollution patterns just before or at MS relapse. Such designs are not optimal to provide causal results, as it answers the question: *Were the levels of air pollution higher prior to relapses?* It implies that the researcher first considered the outcome, that is, relapse occurrence, and then seeks for its environmental causes. This strategy contradicts the principle of classical experimental designs where exposures are assigned randomly prior to measuring the outcome of interest, a method that is the "gold" standard to obtain objective inference on the effects of an intervention.[2] We will follow the steps of Bind and Rubin's[3] causal pipeline to examine the causal question: *Does lowering air pollution levels reduce the risk of relapses?* With this illustration, we aim to engage environmental epidemiologists in: (1) discussing hypothetical interventions that could have resulted in the observed data, (2) verifying the plausibility of the assumptions of the Rubin Causal Model, (3) choosing an adequate data analysis strategy, and (4) interpreting the implications of their results in order to give recommendation for further research or policies.

2 | DATA

2.1 | Multiple sclerosis patients data

The 353 patients in our study are part of the alSacEP network following MS diagnosed patients living in the Alsace region. All patients records were managed with the standardized European Database for Multiple Sclerosis (EDMUS).[27] We focus on the period between 1 January 2000 and 31 December 2009. Two types (1 and 2) with two subtypes (A and B) of patients can be distinguished in the study population (see Figure 1). For Type 1 patients, their relapse history is known from some time post-MS onset, until the end of our study period (Type 1A), or until last patient information (Type 1B). For Type 2 patients, their relapse history is available from MS onset, until the end of our study period (Type 2A), or until last patient information (Type 2B). In epidemiology, Type 1 patients are *prevalent cases*, that is, they were diagnosed with the disease before the study period started. Type 2 patients are *incident cases*, that is, they are newly diagnosed during the study period.

The patients are subject to two forms of MS, the *Relapsing-Remitting* (RR) ($N_{RR} = 296$) and the *Secondary-Progressive* (SP) ($N_{SP} = 57$) form. All the patients started their disease in a Relapsing-Remitting form: the relapses are followed by a remission, that is, a time of recovery with few or no symptoms. But by the time of the study, for some, the disease shifted to a Secondary-Progressive form: the symptoms of the relapses steadily become worse with no remission.[28,29]

Recorded relapses of 109 patients in the alSacEP network may present some doubtful dates, that is, uncertain or completely unknown. Since the outcome of interest of this study is daily relapse occurrence, the analyses in this article are restricted to patients with complete relapses history. Including the patients with inaccurate relapse recording would add an additional source of uncertainty. To take the MS relapse definition[14] into account, for each patient we exclude their 30-day period(s) post-MS onset and post-relapses from the data (see Appendix A for pre-analysis data exclusion details). See Table 1 for the characteristics of the 353 patients included in our subsequent analyses.

2.2 | Relapses in multiple sclerosis are age-, time-, and sex-dependent

Several studies have shown that relapses occurrence are age-, time-, and sex-dependent.[30-32] Relapse rates decrease with time and this decline increases in magnitude with age. Overall, women exhibit a higher relapse rate. We observed a similar relapse rate pattern that is age-, time-, and sex-dependent in our study population (see Web Figure 1), thereby, age, disease history, and sex should be accounted for in our analyses.

2.3 | K-means clustering for patient grouping

The structure of our data implies different timings of disease history for the patients. Overall, Type 1 patients are older, thus at a later stage of their disease progression, and have a longer follow-up period than Type 2 patients during the study period (see Table B1 in Appendix B). However, some Type 1 patients resemble Type 2 patients more (and vice versa) with

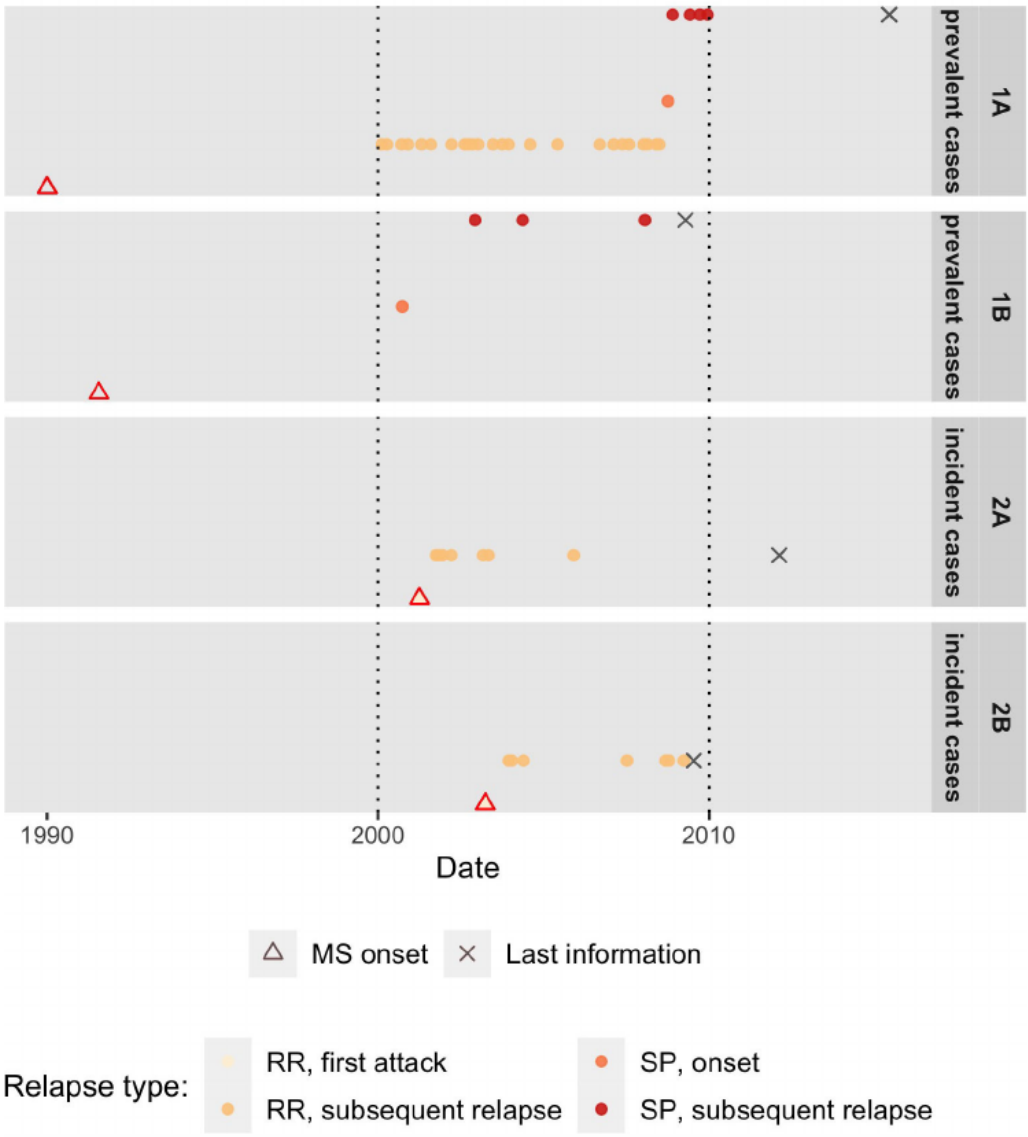

FIGURE 1 Multiple sclerosis patient types with respect to disease onset and relapse records during the study period of 2000-01-01 to 2009-10-01. Type 1—*prevalent cases*: MS onset occurred before the beginning of the study (37%); Type 1A: the patient was followed until the end of the study, Type 1B: the patient's last information was before the end. Type 2—*incident cases*: MS onset occurred after the beginning of the study (between 2000-01-01 and 2009-10-01) (63%); Type 2A: the patient was followed until the end of the study, Type 2B: the patient's last information was before the end [Color figure can be viewed at wileyonlinelibrary.com]

respect to the characteristics: age at study inclusion, disease stage (approximated by the date difference between MS onset and study inclusion), and follow-up duration. For example, a Type 1 patient whose disease onset occurred a year before the beginning of study period at a young age might resemble more Type 2 patients. Therefore, we redefined our patient groups before analyzing our data to provide stratified results taking the timing of disease progression into account. We use the K-means clustering algorithm of Hartigan and Wong[33] provided by the `kmeans` R function[34] to create two clusters that are homogeneous with respect to age at study inclusion, disease stage, and follow-up duration (see Table 1). The groups were homogeneous with respect to the characteristics with two clusters, but not with three or four. In Table 1, we can observe that Cluster 1 patients data is at a later stage of their disease, that is they are older and with a longer onset-study inclusion gap, as compared to Cluster 2 patients (Figure 2).

2.4 | Environmental data

We have meteorological variables for the Alsace region, such as the daily temperature. Air pollution concentrations of particulate matter of 10 μm or smaller in diameter (PM_{10}), and ozone (O_3) were estimated daily at the census block level throughout the study period using the deterministic Atmospheric Dispersion Modeling System (ADMS)-Urban air dispersion model,[35] which included background pollution concentrations, emissions inventories, meteorological data, land use, and surface roughness.

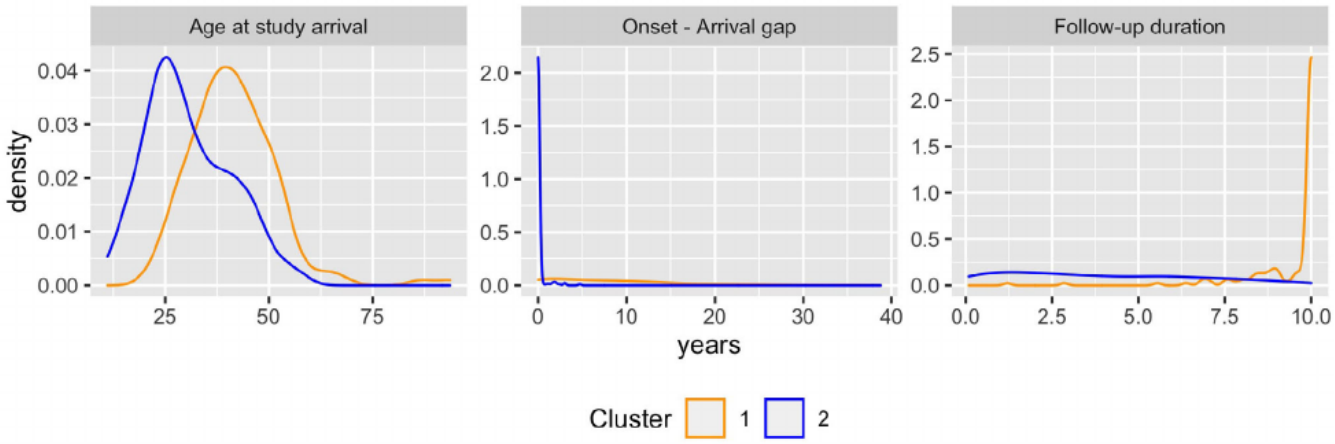

FIGURE 2 Distributions of the patient characteristics for the cluster built with K-means (k = 2) [Color figure can be viewed at wileyonlinelibrary.com]

3 | METHODS

We now present the four stages of the causal pipeline[3] that we use to construct plausible hypothetical randomized experiments to study the air pollution-MS relapse relationship:

- Stage 1: Formulation of a plausible hypothetical intervention decreasing air pollution levels to examine whether it reduces the relapse risk for MS patients.
- Stage 2: Design the hypothetical randomized experiment as if the environmental intervention had been implemented randomly at the census block level.
- Stage 3: Statistical analysis to estimate the relapse risk of MS patients hypothetically randomized to the environmental intervention and test the null hypothesis of no effect of the intervention.
- Stage 4: Interpretation of the estimates obtained from the analysis.

3.1 | Stage 1: Conceptualization of a plausible intervention reducing air pollution levels

3.1.1 | Causal question

We are interested in the causal question: *Does a reduction in PM_{10} levels cause a decrease in relapse occurrence risk for Alsacian MS patients?* However, it is impractical and unethical to expose MS patients to clean air and PM_{10} in a randomized controlled experiment. Therefore, we conceptualize a hypothetical experiment designed to study the effects on MS patients of the following political intervention that reduces the air pollution exposure: *The Alsace region council decides, at the census block-level, to randomly ban some cars to ride, during a few days to keep the average PM_{10} level below or equal to* 15 $\mu g/m^3$. To disentangle the effects of low vs high air pollution levels on the relapse occurrence, the goal is to compare the units under the intervention to units under higher levels of air pollution: average PM_{10} level higher or equal to 25 $\mu g/m^3$. The intervention comparison thresholds are based on the 25th and upper 75th percentiles of the 5 days moving average PM_{10} distribution.

The study population consists of N patients, in S census blocks, followed during T days, where $i = 1, \dots, N$, $s = 1, \dots, S$, and $t = 1, \dots, T$. The objective is to construct a hypothetical experiment that mimics a controlled experiment, in which air pollution exposure could be believed to be randomized. We define the daily census block exposure as the 5 days air pollution moving average. We denote $P_{s,t}$ the 5-day moving average of 24-h-mean PM_{10} in census block s, at day t:

$$P_{s,t} = \frac{1}{5} \sum_{l=1}^{5} PM_{10\ s,t-l}. \quad (1)$$

We chose to calculate the air pollution moving average starting at lag-1 to make sure that the exposure was measured prior to the outcome measured at lag-0. The 5-day moving average is motivated by the results from Jeanjean et al.,[23] who reported positive associations between MS relapse incidence and PM_{10} until lag-5.

The indicator of the intervention vs higher pollution levels for each census block s at day t is

$$W_{s,t} = \begin{cases} 0 & \text{if } P_{s,t} \geq 25\ \mu g/m^3, \\ 1 & \text{if } P_{s,t} \leq 15\ \mu g/m^3. \end{cases} \tag{2}$$

The experimental units are *person days*, that is, patient i in census block s at day t, with intervention indicator: $W_{i,s,t} = W_{s,t}$. In this setting, each unit, has two binary potential outcomes: $Y_{i,s,t}(1)$, the relapse occurrence if $W_{i,s,t} = 1$, and $Y_{i,s,t}(0)$, otherwise:

$$Y_{i,s,t} = \begin{cases} 1 & \text{if a relapse occurred,} \\ 0 & \text{if no relapse occurred.} \end{cases} \tag{3}$$

3.1.2 | Assumptions

To draw causal inferences in a standard setting, the stable unit treatment value assumption (SUTVA) must hold.[36] This assumption incorporates the idea that units do not interfere with one another and that for each unit there is only one single version of each exposure. In the setting of this article, one could argue that some MS patients are still mobile enough to receive hidden versions of the intervention on a day t. But as shown by Jeanjean et al 23, the Alsace region only presents major air pollution contrasts between the census blocks of the main city (Strasbourg) and the surrounding ones. Thus, in this study, we make the assumption that MS patients living in Strasbourg spend most of their time in the city and the patients living in the more rural parts of Alsace do not spend much time in the city.

Another key component of a causal analysis is the assignment mechanism determining which units receive which treatments; in other words, which potential outcomes are observed and which are missing.[36] This study is observational because the functional form of the assignment mechanism is unknown as opposed to a randomized experiment where the assignment mechanism has a known functional form that is controlled by the researcher. Therefore, the researcher has to resort to a design stage to assess the plausibility of an unconfounded assignment mechanism.

3.2 | Stage 2: Design of a reconstructed hypothetical experiment

At the design stage, the aim is to obtain a balanced subset of the observed data for which the assignment to exposure can be assumed to be unconfounded, that is, the exposure assignment is independent of the potential outcomes given the pre-exposure covariates $\mathbf{X}$: $Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = Pr(\mathbf{W}|\mathbf{X})$. Unconfoundedness implies that treated and control groups of units can be fairly compared because they are similar with respect to preexposure covariates.[36] Matching has been a popular method to create treated and control groups that are balanced, that is, exchangeable with respect to their covariates.[37] By creating matched groups we limit the "counfounding" of the exposure-outcome relationship.

As we said earlier, in the definition given by McDonald et al,[14] MS relapses occur for at least 24 hours and start at least 30 days after any previous episode began. This definition leads to no observation of the data on days between t and $t + 30$, if $Y_{i,s,t} = 1$. Therefore, we introduce a clear data indicator, defined as

$$C_{i,s,t} = \begin{cases} 1 & \text{if } \mathbf{Y}_{i,s,t-1:t-30} = 0, \\ 0 & \text{if } \mathbf{Y}_{i,s,t-1:t-30} \neq 0. \end{cases} \tag{4}$$

The days for which $Y^{obs}_{i,s,t} = (Y_{i,s,t}$ $_{i,s,t} = 1)$ can be observed are referred to as *control days*, and the days for which $Y^{obs}_{i,s,t} = (Y_{i,s,t}$ $_{i,s,t} = 1)$ can be observed are referred to as *treated days*.

$(0)|C$

$(1)|C$

3.2.1 | Within-patient pair matching to obtain a balanced subset of the data

Our within-patient matching strategy aims to limit confounding by census block-specific and patient-specific variables. We match our units, *person days*, within patient: patient i in census-block s at time t under $W_{s,t}^{obs} = 1$ with pre-exposure covariates $\mathbf{X}_{i,s,t}$ is matched to himself at $t^\star$, under $W_{s,t^\star}^{obs} = 0$ only if $\mathbf{X}_{i,s,t^\star}$ is "similar" to $X_{i,s,t}$.

For each unit, the vector of covariates is given by $\mathbf{X}_{i,s,t} = (X_{i,s,t}^{(1)}, X_{i,s,t}^{(2)}, X_{i,s,t}^{(3)}, X_{i,s,t}^{(4)})$, where $X_{i,s,t}^{(1)}$ indicates the *number of days elapsed since the MS onset*, $X_{i,s,t}^{(2)}$ the *season*, $X_{i,s,t}^{(3)}$ the *ozone* concentration (in $\mu g/m^3$) at day $t-6$, and $X_{i,s,t}^{(4)}$ the *maximum temperature* (in °C) at day $t-6$. A balanced number of days elapsed since the MS onset ($X_{i,s,t}^{(1)}$) between treated and control days assures that, within-patient, the disease outcomes will be fairly compared at similar points in time during the analysis (stage 3); thereby limiting confounding related to aging and disease progression of the patients. Because of our within-patient matching strategy, we do not have to match patient-specific covariates, such as sex, bodymass index, or smoking status. Matching on $X_{i,s,t}^{(2)}, X_{i,s,t}^{(3)}$ and $X_{i,s,t}^{(4)}$ limits the environmental confounding at the census-block level.

To ensure covariate balance, we only allow a treated unit to be matched with a control unit if the componentwise distances between their covariate vectors are less than some prespecified thresholds $\delta_1, \ldots, \delta_4$. For any pair of covariate vectors $X_{i,s,t}$ and $X_{i,s,t^\star}$, we define the difference between them as

$$\Delta(X_{i,s,t}, X_{i,s,t^\star}) = \begin{cases} 0 & \text{if } |X_{i,s,t}^{(k)} - X_{i,s,t^\star}^{(k)}| < \delta_k \quad \text{for all } k \in \{1,2,3,4\}, \\ +\infty & \text{otherwise} \end{cases} \tag{5}$$

At this stage, the objective to create a balanced data subset for which the plausibility of the "unconfoundedness" assumption is based on a diagnostic of our choice. We choose the thresholds according to the covariates prematching distributions diagnostic plots (see Figure 3: the range and mean of the lag-6 ozone level (in $\mu g/m^3$) are [1, 225] and 63 respectively, and the range and mean of the lag-6 maximum temperature (in °C) are [−10, 39] and 16 respectively). The thresholds are: the absolute difference between the number of days elapsed since the MS onset is less than $\delta_1 = 2$ years, the seasons are identical, that is, $\delta_2 = 0$, the absolute difference in lag-6 ozone level is less than $\delta_3 = 20$ $\mu g/m^3$, and the absolute difference in lag-6 maximum temperature is less than $\delta_4 = 5$°C.

This constrained pair matching can be achieved by using a maximum bipartite matching[38] on a graph such that: (1) there is one node per unit, partitioned into *treated nodes* and *control nodes;* and (2) the edges are pairs of treated and control nodes with covariates $X_{i,s,t}$ and $X_{i,s,t^\star}$; and (3) an edge exists if and only if $\Delta(X_{i,s,t}, X_{i,s,t^\star}) < +\infty$. By construction, using a maximum bipartite matching algorithm on this graph as implemented in the R package `igraph` produces the largest set of matched pairs that satisfy the unit-specific proximity constraints set by our thresholds. The diagnostics for balance show that, the within-patient pair matching algorithm described above was successful in constructing "similar" control (polluted) and treated (clean) days (see the distributions of $X_{i,s,t}$ in both groups in Figure 3). Given the available covariates, our attempt to mimic the randomized intervention from the Alsace city council was successful at creating comparable groups of polluted vs less polluted days.

3.3 | Stage 3: Analysis of the hypothetical experiment

In this illustration, we follow a Fisherian analysis approach and perform hypothesis testing with a powerful test statistic comparing relapse occurrence of units subject to an intervention for air pollution reduction to units subject to higher levels of air pollution.[39] We do not attempt to provide an estimate of (and uncertainty around) an estimand to avoid relying on assumptions such as the additivity of the treatment effects, asymptotic arguments, or an imputation model, which may be the case when drawing Neymanian or Bayesian inferences.

3.3.1 | Sharp null hypothesis

The sharp null hypothesis, stating that for each unit the intervention (exposure) has no effect on "clear days" (ie, $C_i = 1$), can be formally expressed as:

$$\forall i,s,t \quad H_0 : (Y_{i,s,t}^{(0)|C_{i,s,t}=1} = Y_{i,s,t}^{(1)|C_{i,s,t}=1}). \tag{6}$$

The plausibility of this sharp null hypothesis can be assessed by using a randomization test.

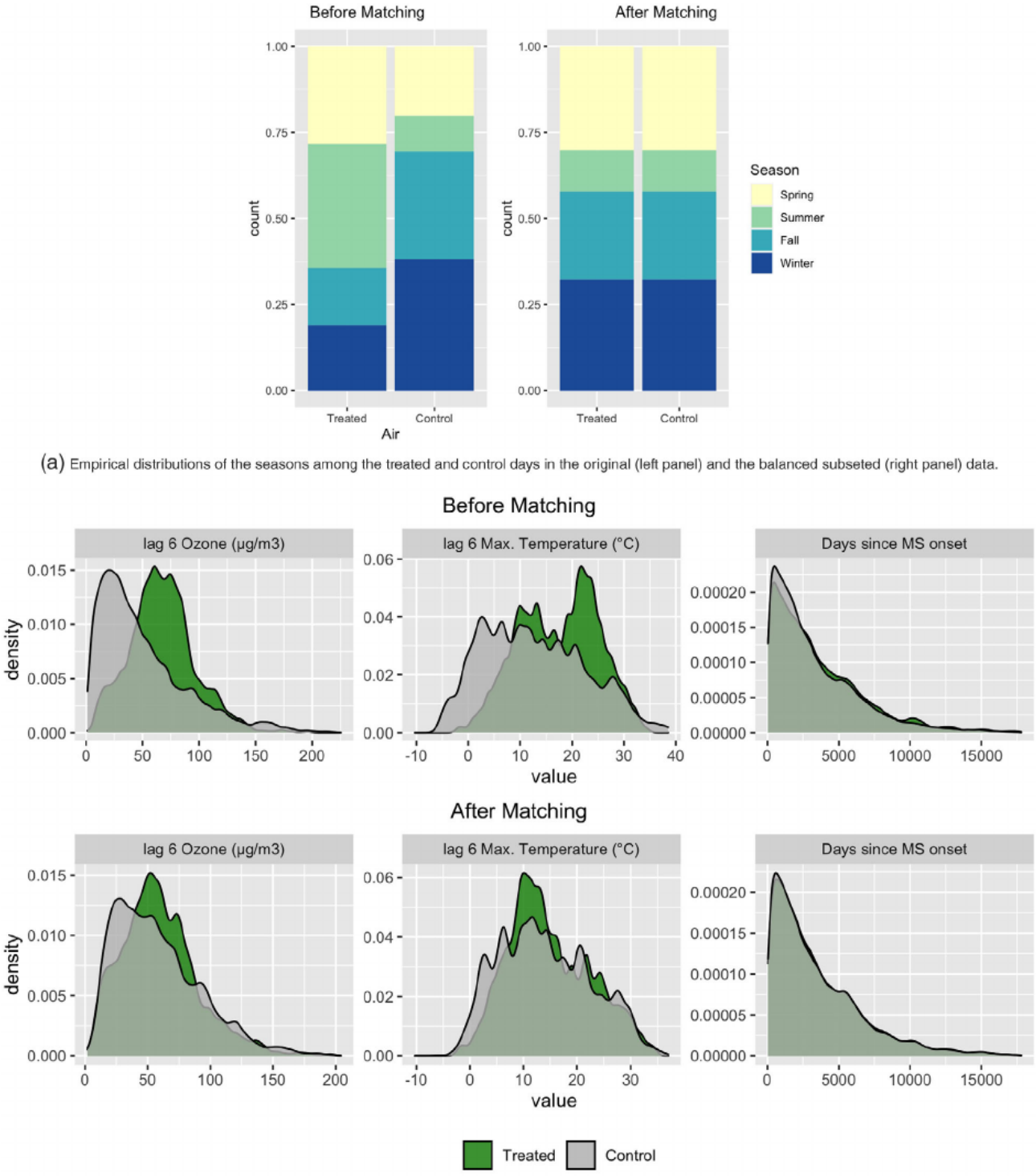

FIGURE 3 Empirical distributions of the preexposure covariates before and after the design stage [Color figure can be viewed at wileyonlinelibrary.com]

3.3.2 | Choice of test statistic

To assess the null hypothesis of no effect of the intervention, we first compute the observed value of a test statistic. We propose to use the β_1 estimated by a logistic mixed effect model as detailed below. Brillinger et al[40] were pioneers to use the coefficient of a model for the statistic for the Fisher randomization test. At this stage, to achieve larger bias reductions, frequentist regression models can be used to remove residual confounding that was not accounted for, during the design stage.[10,41] Because the outcome of interest is a binary response whose mean is conditional on the patient i, we consider a

logistic mixed effects model for $Y_{i,s,t}$.[42] The log odds of $Y_{i,s,t}$ depend on fixed and random effects via the following linear predictor:

$$log\left(\frac{Pr(Y_{i,s,t}=1|C_{i,s,t}=1,b_i)}{Pr(Y_{i,s,t}=0|C_{i,s,t}=1,b_i)}\right)=W_{s,t}\beta_1+A_{i,s,t}\beta_2+b_i\,, \tag{7}$$

where $A_{i,s,t}$ is the age of patient i at MS onset.

That is, the conditional mean of $Y_{i,s,t}$ is related to the linear predictor by a logit link function and has randomly varying intercepts b_i taking the natural heterogeneity of units' propensity to have a relapse. We adjust for $A_{i,s,t}$, *age at onset*, because it has been observed that the younger the patient at disease onset, the higher its relapse rate.[32,43,44] In Equation (7), the β_1 coefficient corresponds to the change in log odds of a relapse when a patient is subject to the air pollution reduction intervention vs higher pollution levels. We estimate β_1 using the `glmer` function of the R package `lme4`.[45]

3.3.3 | Randomization-based inference

Here, we choose not to rely on asymptotic arguments, but instead take a Fisherian perspective (ie, randomization-based inference).[39,46] Assuming H_0, the goal is to approximate the null randomization distribution of β_1, β_1^{null} by computing the values of the test statistic for 1000 possible exposure assignments. Because the number of assignments is very large, we calculate an approximate P-value, that is, the proportion of computed test statistics that are as large or larger than the observed test statistic: $\frac{1}{1{,}000}\sum_{r=1}^{1{,}000}\mathbb{1}_{|\beta_{1,r}^{null}|\geq|\beta_1^{obs}|}$, where $\mathbb{1}_{|\beta_{1,r}^{null}|\geq|\beta_1^{obs}|}=1$ when $|\beta_{1,r}^{null}|\geq|\beta_1^{obs}|$ and, 0 otherwise. A small P-value shows that the observed test statistic is a rare event when the null hypothesis is true. When units have varying probabilities of being treated, the analysis of experiments, even when hypothetical, should reflect their design.[2,3,11,47] In our example, patients living in the same census block have the same intervention exposure. We consider two hypothetical intervention assignment mechanisms operating at the censusblocklevel: *Every day t, in each census block s, the Alsace city council decides to impose the air pollution reduction intervention using a (1) completely randomized and (2) temporally correlated assignment mechanism*. Therefore, we generated 1000 exposure assignments at every day t, in each census block s:

1. by tossing a coin with probability $Pr(W_{s,t}=1)=1/2$, and
2. by generating $W_{s,t}$ with auto-correlation: $Cor(W_{s,t},W_{s,t-1})=0.5$, where 0.5 corresponds to the air pollution correlation of adjacent days in the data.

3.4 | Stage 4: Interpretations of the results

If the null hypothesis of no difference in MS relapses between the matched groups of treated and control units is rejected, that difference warrants further scrutiny to assess whether it can be attributed to the different air pollution levels, assuming the assignment "unconfoundness" assumption holds. We can then report that the relapse risk of MS patients was or was not reduced by the introduction of the intervention to reduce air pollution levels in the Alsace region. It is important to note that interpretation should be restricted to units that remain in the finite sample after matching (see their detailed characteristics in Table 1 and Figure 3). The data do not provide direct information for "unmatched" units. Cautiousness regarding extrapolation to units with covariate values beyond values observed in the balanced subset of the data is necessary. The results of our analyses and associated discussion are presented next.

4 | RESULTS

Our balanced subset of the data was analyzed as a whole, and within patient Clusters 1 and 2 to assure we study patients that are in similar phases of their disease history. Recall that Cluster 1 patients are older, thus at a later stage of their disease progression, and have a longer follow-up period than Cluster 2 patients (see Table 1). Accordingly, we anticipate Cluster 1 patients to develop fewer relapses than Cluster 2 patients, regardless of their environmental exposure. In the matched population, we estimated the log odds of a relapse after the patients are subject to a hypothetical intervention decreasing the air pollution levels vs higher pollution levels. These estimates and their associated approximate Fisherian P-values,

TABLE 2 Primary results

		Control days	Treated days	Estimate	P-value$_{CR}$	P-value$_{TC}$
	O	185 942	160 186			
	B	89 410	89 410	−0.12	*.341*	*.485*
C1	O	118 350	99 677			
	B	57 969	57 969	−0.04	*.842*	*.862*
C2	O	67 592	60 509			
	B	31 441	31 441	−0.23	*.227*	*.384*

Note: Estimates and approximate Fisherian P-values calculated in the balanced data subset (B vs original (O)) for the overall sample and stratified by patient clusters (C1 and C2). We consider the completely randomized (CR) and temporally correlated (TC) assignment mechanisms.

based on 1000 draws of the permuted treatment assignment are presented in Table 2. We also present the secondary results stratified by sex and MS form, determined by the last patient information (see Table C1 in Appendix C).

The sharp null hypothesis of no effect of the intervention lowering the levels of air pollution in the overall study population is not rejected (P-value$_{CR} = .341$ and P-value$_{TC} = .485$, see Table 2). However, in the block of female patients with a relapsing-remitting MS form there is an indication that the observed intervention effect could be a rare event under the null hypothesis (P-value$_{CR} = .038$ and P-value$_{TC} = .160$, see Table C1 in Appendix C). To assess the significance of this secondary result rigorously, another study primarily focusing on this subpopulation should be conducted.

5 | DISCUSSION

We have illustrated Bind and Rubin's[3] causal inference pipeline with complex environmental health data. Standard epidemiological approaches analyze the observed data by directly regressing an observed outcome on an exposure and confounding covariates. Instead, before analyzing the exposure-outcome (pollution-MS) relationship, we constructed a balanced data set in such a way that it could have plausibly come from an intervention that we conceptualized. The objective of such approach is to borrow the appealing insights of randomized control trials in observational studies.

During the design stage, the outcome variable is ignored and only pre-exposure covariates are considered. The chosen balanced data is a subsample of units that can be used to estimate the effects of an exposure in potentially susceptible populations. This advantage is particularly interesting for epidemiological studies because it facilitates the study of non-modifiable risk factors (eg, race, age, sex). Omitting the outcome data until the analysis avoids "model cherry-picking" because the effect of the intervention is estimated once, only after the design stage is successful. Nonetheless, at the design stage, we can only consider the observed preexposure variables but the assignment mechanism could depend on unobserved preexposure variables. In such case, it is recommended to consider sensitivity analyses of how the Fisher P-value would change had the assignment mechanism been plausibly different, as suggested by Rosenbaum[48] and further discussed by Bind and Rubin.[49] Subject-matter knowledge on air pollution exposure assessment should guide the plausible range of "sensitivity". P-values and the reason why they could deviate from the P-value calculated with the assumed hypothetical assignment mechanism.

Results interpretation are more transparent than with standard approaches. The assumed assignment mechanism and underlying assumptions have to be clearly stated to obtain meaningful P-values. Standard approaches usually make strong assumptions (eg, linearity), whose discussions are often neglected. Solely adjusting for confounders by including them in a regression function, without a design stage, can be unreliable, especially when the pre-exposure covariates distributions of the treated and control units are not similar. Cochran and Rubin,[41] Heckman et al,[50] and Rubin[51] have shown that regression models can estimate biased treatment effects when the true relationship between the covariates and the outcome is not modeled accurately. Nonetheless, the temporal structure of our study could question the plausibility of the "no hidden version of the treatment" component of the SUTVA. One could argue that the small P-values we reported are due to air pollution exposure that happened prior to t. Therefore, with the same analysis method, we verified (in the female patients block) that the null hypothesis of no effect of the intervention was not rejected for $W_{s,t-1}$

(*estimate* = −0.16, *P*-value = .223) and $W_{s,t-2}$ (*estimate* = −0.03, *P*-value = .843), the intervention indicators summarizing the 5-day moving average of 24-h-mean PM_{10} in census block s, at day $t-1$ and $t-2$ respectively (see Equations (1) and (2) in the Methods). Concerning the interference component of the SUTVA, omitting the 30 days post-relapse reassures the absence of person days interference impacting the observed outcome. Another question that could arise due to the temporality in our data, is whether the assignment mechanism depends on historical information of covariates and exposures. We assure it is not the case by verifying that the covariates and exposures are balanced post-matching until lag 10 (see Web Figure 2).

In contrast to other studies interested in the effect of air pollution exposures on health outcomes, this study does not provide any estimation of an exposure-response curve. Instead, we chose to estimate the effect of a single intervention and provide results that can directly contribute to policy recommendations. The agency in charge of monitoring the Alsace region air quality, *Atmo Grand Est*, informs and warns the citizens, medias, and local governments on the air pollution levels. For instance, during the Summer 2019 heat wave, the local government imposed a reduction of automobile speed of 20 km/h on the highway. These interesting interventions are intended to prevent the harmful effects of high pollution episodes. We believe that research that is intervention oriented, as conducted in this study, should help policy makers in better tailoring their intervention policies to prevent adverse health effects of environmental exposures. Also, until now no environmental epidemiologists analyzed the air pollution-MS relationship with causal inference methods, so a first step to make advances in the field is to assess, by comparing low vs high air pollution exposure, such sharp null hypothesis, that is, whether air pollution has any effect on the units of our study. If so, a natural next step would be to work with a data set adequate for balancing covariates along different doses of the exposure such as suggested by Wu et al[9] and estimate a causal dose response to protect populations at risk.

The null hypothesis of no effect of air pollution reduction intervention is not rejected in the overall study population, which differs from previous studies,[17-23] and highlights the statistical conclusion differences between studies using causal inference methods vs directly modeling the observed data (eg, using regressions). The secondary analyses indicate an effect of the intervention that is worthy of attention in the subgroup of women with relapsing-remitting MS; such question was not examined in previous studies. The effects of air pollution may be different between men and women.[52,53] It has to be reminded that this subgroup of women was not the primary focus of our study, this result has to be confirmed in a study designed for this subgroup.

A limitation of our study is that we had to focus on the patients for whom we had a complete disease history and omit the patients whose relapses were recorded on a doubtful date. Ideally, we should have imputed the dates of these relapses by following a multiple imputation procedure for outcome data as suggested by Little and Rubin.[54] However, the causes of MS relapses, a rare outcome, remain unknown, which makes their timing nearly impossible to predict accurately. This issue motivates why we decided to analyze a complete-case subset of MS patients. Furthermore, we considered only one pollutant, PM_{10}, which constitutes another limitation of our study. The environmental epidemiology literature suggests that a pollutant mixtures may be more relevant to study. Our illustration could be extended: (1) to the estimation of an exposure-response curve to protect populations at risk, (2) to the estimation of the effects of an intervention involving multiple exposures on the risk of MS relapse, and (3) to study effects of air pollution decrease interventions on other health outcomes, such as stroke or asthma exacerbation.[55-57]

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the Office of the Director, National Institutes of Health under Award Number DP5OD021412 and the John Harvard Distinguished Science Fellows Program within the FAS Division of Science of Harvard University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Prof D. B. Rubin for many helpful discussions and suggestions, Dr Stéphane Shao for his help with the implementation of the matching algorithm, Prof de Sèze and Dr. Ongagna for sharing the data set, Mrs Berthe and Mr Senger for their help during the data collection and quality control, all the neurologists contributing to the alSacEP network, and Dr Bard, Dr Roux, and Dr. Jeanjean for making this project possible.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Research data cannot be shared for the protection of patients identity and associated medical records.

ORCID

Alice J. Sommer https://orcid.org/0000-0003-0989-4103
Marie-Abèle C. Bind https://orcid.org/0000-0002-0422-6651

REFERENCES

1. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701.
2. Rubin DB. For objective causal inference design trumps analysis. *Ann App Stat*. 2008;2(3):808-840.
3. Bind M-AC, Rubin DB. Bridging observational studies and randomized experiments by embedding the former in the latter. *Stat Methods Med Res*. 2019;28(7):1958-1978.
4. Sommer AJ, Lee M, Bind MAC. Comparing apples to apples: an environmental criminology analysis of the effects of heat and rain on violent crimes in Boston. *Palgrave Commun*. 2018;4(1):1-10.
5. Zigler CM, Dominici F. Point: clarifying policy evidence with potential-outcomes thinking—beyond exposure-response estimation in air pollution epidemiology. *Am J Epidemiol*. 2014;180(12):1133-1140.
6. Baccini M, Mattei A, Mealli F, Bertazzi PA, Carugno M. Assessing the short term impact of air pollution on mortality: a matching approach. *Environ Health A Global Access Sci Source*. 2017;16(1):7.
7. Dominici F, Zigler C. Best practices for gauging evidence of causality in air pollution epidemiology. *Am J Epidemiol*. 2017;186(12):1303-1309.
8. Bind MA. Causal modeling in environmental health. *Annu Rev Public Health*. 2019;40(1):23-43.
9. Wu X, Braun D, Schwartz J, Kioumourtzoglou MA, Dominici F. Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. *Sci Adv*. 2020;6(29):eaba5692.
10. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*. 1973;29(1):185-203.
11. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;26(1):20-36.
12. Xu X, Ha SU, Basnet R. A review of epidemiological research on adverse neurological effects of exposure to ambient air pollution. *Front Publ Health*. 2016;4:157.
13. Mousavi SE, Heydarpour P, Reis J, Amiri M, Sahraian MA. Multiple sclerosis and air pollution exposure: mechanisms toward brain autoimmunity. *Med Hypotheses*. 2017;100(C):23-30.
14. McDonald WI, Compston A, Edan G, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. *Ann Neurol*. 2001;50(1):121-127.
15. Ascherio A, Munger KL, White R, et al. Vitamin D as an early predictor of multiple sclerosis activity and progression. *JAMA Neurol*. 2014;71(3):306-314.
16. Belbasis L, Bellou V, Evangelou E, Ioannidis JPA, Tzoulaki I. Environmental risk factors and multiple sclerosis: an umbrella review of systematic reviews and meta-analyses. *Lancet Neurol*. 2015;14(3):263-273.
17. Oikonen M, Laaksonen M, Laippala P, et al. Ambient air quality and occurrence of multiple sclerosis relapse. *Neuroepidemiology*. 2003;22:95-99.
18. Gregory AC, Shendell DG, Okosun IS, Gieseker KE. Multiple sclerosis disease distribution and potential impact of environmental air pollutants in Georgia. *Sci Total Environ*. 2008;396(1):42-51.
19. Heydarpour P, Amini H, Khoshkish S, Seidkhani H, Sahraian MA, Yunesian M. Potential impact of air pollution on multiple sclerosis in Tehran. *Iran Neuroepidemiol*. 2014;43:233-238.
20. Angelici L, Piola M, Cavalleri T, et al. Effects of particulate matter exposure on multiple sclerosis hospital admission in Lombardy region, Italy. *Environmental Research*. 2016;145(Supplement C):68-73.
21. Roux J, Bard D, Le Pabic E, et al. Air pollution by particulate matter (PM10) may trigger multiple sclerosis relapses. *Environ Res*. 2017;156:404-410.
22. Bergamaschi R, Cortese A, Pichiecchio A, et al. Air pollution is associated to the multiple sclerosis inflammatory activity as measured by brain MRI. *Multiple Sclerosis*. England: Houndmills, Basingstoke; 2017.
23. Jeanjean M, Bind M-A, Roux J, et al. Ozone, NO2 and PM10 are associated with the occurrence of multiple sclerosis relapses. Evidence from seasonal multi-pollutant analyses. *Environ Res*. 2018;163:43-52.
24. Palacios N, Munger KL, Fitzgerald KC, et al. Exposure to particulate matter air pollution and risk of multiple sclerosis in two large cohorts of US nurses. *Environ Int*. 2017;109:64-72.
25. Chen H, Kwong JC, Copes R, et al. Living near major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis: a population-based cohort study. *The Lancet*. 2017;389(10070):718-726.
26. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*. 1991;133(2):144–153.
27. Confavreux C, Compston DA, Hommes OR, McDonald WI, Thompson AJ. EDMUS a European database for multiple sclerosis. *J Neurol Neurosurg Psychiatry*. 1992;55(8):671–676.
28. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. *Neurology*. 1996;46(4):907-911.
29. Lublin FD, Reingold SC, Cohen JA, et al. Defining the clinical course of multiple sclerosis. *Neurology*. 2014;83(3):278-286.
30. Weinshenker BG, Bass B, Rice GPA, et al. The natural hisory of multiple sclerosis: a geographically based study: I clinical course and disability. *Brain*. 1989;112(1):133-146.

31. Confavreux C, Vukusic S, Moreau T, Adeleine P. Relapses and progression of disability in multiple sclerosis. *N Engl J Med*. 2000;343(20):1430-1438.
32. Tremlett H, Zhao Y, Joseph J, Devonshire V. Relapses in multiple sclerosis are age- and time-dependent. *J Neurol Neurosurg Psychiatry*. 2008;79(12):1368–1374.
33. Hartigan JA, Wong MAA. K-means clustering algorithm. *J Royal Stat Soc Ser C (Appl Stat)*. 1979;28(1):100-108.
34. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013.
35. Carruthers D, Edmunds HA, Lester AE, McHugh CA, Singles RJ. Use and validation of ADMS-Urban in contrasting urban and industrial locations. *Int J Environ Pollut*. 2000;14:364-374.
36. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press; 2015.
37. Rubin DB. *Matched Sampling for Causal Effects*. New York, NY: Cambridge University Press; 2006.
38. Micali S, Vazirani VV. An Algoithm for Finding Maximum Matching in General Graphs. *SFCS '80*. Washington, DC: IEEE Computer Society; 1980:17-27.
39. Fisher RA. *The Design of Experiments*. Edinburgh: Oliver and Boyd; 1935.
40. Brillinger DR, Jones LV, Tukey John W. *The Role of Statistics in Weather Resources Management*. Vol 25. Washington, DC: U.S. Government Printing Office; 1978.
41. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhya Indian J Stat Ser A (1961–2002)*. 1973;35(4):417-446.
42. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. 2nd ed. Hoboken, NJ: Wiley; 2011.
43. Confavreux C, Vukusic S. Age at disability milestones in multiple sclerosis. *Brain*. 2006;129(3):595-605.
44. Leray E, Yaouanq J, Le Page E, et al. Evidence for a two-stage disability progression in multiple sclerosis. *Brain*. 2010;133(7):1900-1913.
45. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1-48.
46. Rubin DB. Randomization analysis of experimental data: the fisher randomization test comment. *J Am Stat Assoc*. 1980;75(371):591-593.
47. Branson Z, Bind MA. Randomization-based inference for Bernoulli trial experiments and implications for observational studies. *Stat Methods Med Res*. 2019;28(5):1378–1398.
48. Rosenbaum PR. *Design of Observational Studies*. 1st ed. New York, NY: Springer-Verlag; 2010.
49. Bind M-AC, Rubin DB. When possible, report a fisher-exact P value and display its underlying null randomization distribution. *Proc Natl Acad Sci*. 2020;117(32):19151-19158. https://doi.org/10.1073/pnas.1915454117.
50. Heckman JJ, Ichimura H, Todd P. Matching as an econometric evaluation estimator. *Rev Econ Stud*. 1998;65:261-294.
51. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcome Res Methodol*. 2001;2(3):169-188.
52. Clougherty J. A growing role for gender analysis in air pollution epidemiology. *Environ Health Perspect*. 2010;118(2):167-176.
53. Oiamo T, Luginaah I. Extricating sex and gender in air pollution research: a community-based study on cardinal symptoms of exposure. *Int J Environ Res Public Health*. 2013;10(9):3801-3817.
54. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. 2nd ed. Hoboken, NJ: Wiley; 2002.
55. Khreis H, Kelly C, Tate J, Parslow R, Lucas K, Nieuwenhuijsen M. Exposure to traffic-related air pollution and risk of development of childhood asthma: a systematic review and meta-analysis. *Environ Int*. 2017;100:1-31.
56. Yuan S, Wang J, Jiang Q, et al. Long-term exposure to PM2.5 and stroke: a systematic review and meta-analysis of cohort studies. *Environ Res*. 2019;177:108587.
57. Graber M, Mohr S, Baptiste L, et al. Air pollution and stroke. a new modifiable risk factor is in the air. *Rev Neurol*. 2019;175(10):619–624.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Sommer AJ, Leray E, Lee Y, Bind M-AC. Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship. *Statistics in Medicine*. 2020;1–15. https://doi.org/10.1002/sim.8843

APPENDIX A. DATA PREPROCESSING STEPS

1. All the patients in the alSacEP network for whom data are available during study period: from 1 January 2000 to 31 December 2009.

2. We removed the patients who recorded relapses only on doubtful dates. Patients who only have their disease onset date unclear were kept because we are analyzing relapses post-onset for all patients.
3. For each patient, we observe their daily data as from 30 days post-MS onset until their last available information. Also, we do not observe patient data 30 days post-relapse because according to McDonald et al's[14] definition, relapses are only recorded when they occurred at least 30 days after the last relapse.
4. We observe days where the 5 days average PM_{10} level are smaller or equal to 15 $\mu g/m^3$, and bigger or equal to 25 $\mu g/m^3$.
5. For six patients, because their number of observed days is small (between 9 and 33 days), no match according to our matching criteria has been found (Table A1).

TABLE A1 Data preprocessing

	Step	$N_{patients}$
1	Original data	473
2	Complete disease history	364
3	30 days post-MS onset to day of last information	355
4	Dichotomization for binary exposure	353
5	Person days matching	347

APPENDIX B. SAMPLE CHARACTERISTICS BY TYPE

TABLE B1 Multiple sclerosis patient characteristics (in years) by Type

Mean in years (SD)	Type 1 (n = 132)	Type 2 (n = 221)
Onset–Inclusion gap	9	0
	(8)	(0)
Age at MS clinical onset	29	32
	(10)	(11)
Age at study inclusion	39	32
	(12)	(11)
Follow-up since inclusion	9	4
	(1)	(3)

Note: Type 1—*prevalent cases*: MS onset occurred before the beginning of the study. Type 2—*incident cases*: MS onset occurred after the beginning of the study (between 2000-01-01 and 2009-10-01).

APPENDIX C. RESULTS

TABLE C1 Secondary analysis

		Control days (n)	Treated days (n)	Estimate	P-value$_{CR}$	P-value$_{TC}$
M	O	51 118	42 977			
	B	24 643	24 643	0.17	*.518*	*.602*
RR	O	38 406	32 087			
	B	18 403	18 403	0.23	*.482*	*.541*
SP	O	12 712	10 890			
	B	6240	6240	0.00	*1.000*	*1.000*
F	O	134 824	117 209			
	B	64 767	64 767	−0.20	*.162*	*.325*
RR	O	105 713	93 504			
	B	50 615	50 615	−0.32	*.038*	*.160*
SP	O	29 111	23 705			
	B	14 152	14 152	0.19	*.517*	*.647*
Cluster 1						
		Control days (n)	**Treated days (n)**	**Estimate**	P-value$_{CR}$	P-value$_{TC}$
M	O	32 391	25 273			
	B	15 312	15 312	0.21	*.515*	*.669*
F	O	85 959	74 404			
	B	42 657	42 657	−0.11	*.529*	*.668*
Cluster 2						
		Control days (n)	**Treated days (n)**	**Estimate**	P-value$_{B,CR}$	P-value$_{B,TC}$
M	O	18 727	17 704			
	B	9331	9331	−0.09	*.809*	*.785*
F	O	48 865	42 805			
	B	22 110	22 110	−0.31	*.137*	*.318*

Note: Estimates and approximate Fisherian P-values calculated in the balanced subset (B vs original (O)). The results are stratified by MS form (RR and SP) within sex (M and F), and by sex within patient clusters (C1 and C2). We consider the completely randomized (CR) and temporally correlated (TC) assignment mechanisms.