

Aus dem Helmholtz Zentrum München,
Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH)
Institut für Epidemiologie und Environmental Health Center (EHC)
Aufsichtsratsvorsitzende: MinDir'in Prof. Dr. Veronika von Messling

Randomization-based causal inference for Environmental Epidemiology



Dissertation
zum Erwerb des Doktorgrades der Naturwissenschaften
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

vorgelegt von

Alice Joséphine Sommer

aus

Brüssel (Belgien)

2022

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Betreuerin: Prof. Dr. Annette Peters

Zweitgutachterin: Prof. Dr. Anne-Laure Boulesteix

Dekan: Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung: 14.03.2023



LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Dean's Office
Faculty of Medicine



Affidavit

Sommer, Alice

Street

Zip code, town
Germany

Country

I hereby declare, that the submitted thesis entitled

Randomization-based causal inference for Environmental Epidemiology

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

Frankfurt am Main, 30.03.2023

Place, date

Alice Sommer

Signature doctoral candidate

To my parents,

Table of contents

List of publications	i
Contribution to publications	ii
Summary	iii
Zusammenfassung	iv
1 Introduction	1
1 Causal inference in environmental epidemiology	1
2 Two studies of adverse health effects of air pollution	5
3 Outlook	6
2 Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship	14
3 A randomization-based causal inference framework for uncovering environmental exposure effects on human gut microbiota	36
Acknowledgements	98

List of publications

A. J. Sommer, E. Leray, Y. Lee, M.-A. C. Bind. Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship. *Statistics in Medicine*, 40(6):1321-1335, 2021

A. J. Sommer, A. Peters, M. Rommel, J. Cyrus, H. Grallert, D. Haller, C. L. Müller, M.-A. C. Bind. A randomization-based causal inference framework for uncovering environmental exposure effects on human gut microbiota. *PLOS Computational Biology*, 18(5):e1010044, 2022

Contribution to publications

CHAPTER 2. AJS and MACB conceptualized and designed the research. AJS analyzed the data and produced all the figures. YL implemented the temporal randomization test. EL prepared the datasets and supported the discussion. AJS wrote the first draft of the paper and thoroughly discussed and edited it with MACB. All authors, read, revised, and approved the published version of the manuscript.

CHAPTER 3. AJS, MACB, AP, and CLM conceptualized the research. AJS analyzed the data and produced all the figures. AJS, MACB, and CLM developed the methods. AJS wrote the first draft of the paper and thoroughly discussed and edited it with MACB and CLM. AJS, MACB, AP, CLM, DH, MT, HG, and JC reviewed and edited it. MACB and AP acquired funding. All authors, read, revised, and approved the published version of the manuscript.

Summary

Environmental epidemiology is the scientific field studying the effects of environmental exposures on human health outcomes, such as the effects of air pollution exposure on cardiovascular diseases. This research proposes statistical methods that the field of environmental epidemiology could benefit from. Most often, when addressing environmental epidemiology questions, the data collected for statistical analyses come from observational studies. In these studies, the assignment of participants into an exposed versus a control group is outside the control of the investigator, as opposed to individual assignments in randomized controlled experiments. The lack of experimental design, with a randomly assigned exposure, typically prevents direct comparisons of exposed and control groups because of differing background covariates distributions. Due to this design barrier, environmental epidemiologists are mainly able to estimate associations between environmental exposures and health outcomes instead of focus on causal effects that can only be estimated when the conditions of a randomized experiment hold.

In the 1970s, the Rubin Causal Model (RCM) was developed and can help remedy the constraints of observational studies. This model has been extensively used in economics; Nobel Prizes were awarded in 2021 to professors whose scientific advances mainly rely on the RCM. Nonetheless, the appealing properties of the RCM remain to be thoroughly introduced to the field of environmental epidemiology by tackling valuable research questions. Therefore, this thesis illustrates how a multi-staged pipeline relying on the RCM can be used in environmental epidemiology to construct and analyze hypothetical randomized experiments with two studies:

1. by investigating whether an air pollution reduction intervention could have an effect on the risk of multiple sclerosis relapses, and
2. by exploring the effects of two inhaled environmental exposures previously hypothesized to be linked with the gut microbiome: air pollution exposure and cigarette smoking.

Zusammenfassung

Die vorliegende Dissertation stellt statistische Methoden vor, welche die Arbeit von Umweltepidemiologen unterstützt. Die Umweltepidemiologie ist jenes wissenschaftliche Gebiet, welches sich mit den Auswirkungen der Umwelt auf die menschliche Gesundheit befasst, z.B. mit den Auswirkungen der Luftverschmutzung auf Herz-Kreislauf-Erkrankungen. Wenn es um Fragen der Umweltepidemiologie geht, werden die Daten für statistische Analysen meist auf Beobachtungsstudien basiert. Bei diesen Studien liegt die Einteilung der Probanden in eine Treatment- und eine Kontrollgruppe außerhalb des Einflussbereichs des Forschers, im Gegensatz zu randomisierten kontrollierten Experimenten. Das Fehlen eines Versuchsplans mit einer randomisierten Zuweisung für die Testung erschwert daher den Vergleich zwischen Treatment- und Kontrollgruppen. Diese Barriere führt dazu, dass Umweltepidemiologen hauptsächlich Beziehungen zwischen Umweltexpositionen und gesundheitlichen Folgen abschätzen, anstatt sich auf kausale Effekte zu konzentrieren, die nur erarbeitet werden können, wenn die Bedingungen eines randomisierten Experiments gegeben sind.

In den 70er Jahren wurde das Rubin Causal Modell (RCM) entwickelt, das die Limitationen von Beobachtungsstudien lösen kann. Dieses Modell wurde im Bereich der Wirtschaftswissenschaften bereits ausgiebig genutzt; ein Nobelpreis wurde im Jahr 2021 an Professoren verliehen, deren wissenschaftliche Fortschritte hauptsächlich auf dem RCM beruhen. Dennoch müssen die aussagekräftigen Eigenschaften des RCM in den Bereich der Umweltepidemiologie eingeführt werden, indem nützliche Forschungsfragen angegangen werden. Daher wird in dieser Dissertation anhand von zwei Studien gezeigt, wie eine mehrstufige Pipeline, die sich auf das RCM stützt, helfen kann, hypothetische randomisierte Experimente im Bereich der Umweltepidemiologie zu rekonstruieren und zu analysieren:

1. durch die Untersuchung, ob eine Intervention zur Verringerung der Luftverschmutzung das Risiko eines Multiple-Sklerose-Schubs verringert; und
2. durch die Untersuchung der Auswirkungen von zwei inhalativen Umweltexpositionen, (Luftverschmutzung und Zigarettenrauchen) auf das Darmmikrobiom.

Chapter 1

Introduction

Environmental epidemiology studies the effects of environmental exposures on health outcomes mainly with observational data. To analyze such type of data, regression models are used, leading the field to make conclusions based on associations even though causal effects are sought. Therefore, in this thesis, we illustrate how to rigorously conduct environmental epidemiology with a causal inference framework based on two applications investigating: (i) air pollution effects on multiple sclerosis relapses risk [Sommer et al., 2021], and (ii) environment-host microbiome relationships [Sommer et al., 2022]. In both studies, we navigate scientists through a framework testing plausible sharp null hypotheses of no adverse effect of an environmental intervention. We chose a Fisherian approach and to test sharp null hypotheses because the relationships we investigate have not been examined with causal inference methods yet. Testing whether the exposures have any effect on the units of our study is a good starting point before estimating causal effects and uncertainty around them (e.g., using a Bayesian causal model), as well as before studying populations with broader participant characteristics (e.g., ethnicity or race). The framework relies on ideas developed in the 70s [Cochran and Rubin, 1973, Rubin, 1973, 1974, 1976] and the Rubin Causal Model (RCM) [Holland, 1986, Imbens and Rubin, 2015]. This method was recently made explicit to the environmental epidemiology field by Bind and Rubin [2017], who present how to analyze observational data by constructing the ideal conditions of randomized experiments, the “gold standard” to draw objective causal inferences on the effects of an intervention [Rubin, 2008].

1 Causal inference in environmental epidemiology

Readers of environmental epidemiology journals are interested in finding evidence of the effects of environmental exposures, or interventions on health outcomes. However, environmental epidemiologists usually estimate associations, by regressing an observed outcome on a function of observed covariates using observational data. Regression use in the field is spread at least for two reasons. First, it is often unethical to randomize humans to possibly harmful environmental exposures, so observational data are mainly collected. Thus,

confounding adjustments arise post-data collection instead of by design in randomized experiments, where the design is thought to answer a specific cause and effect question. Second, the rise of statistical software on personalized desktops at universities in the 1980s triggered preferences for rapid answers from regression models. This was when investigators transitioned away from planned statistical analyses, thereby omitting not only the formal definition of a causal estimand, but also the comparison of comparable groups of treated and control units [Bind, 2019]. Therefore, causal modelling techniques to analyze observational data are currently being introduced or re-introduced to the field to stimulate reporting results on the health effects of environmental interventions instead of mere association reporting.

For decades, environmental epidemiology has suffered from the fact that association is not causation and therefore the field recently tackled the confounding issue with methods such as Directed Acyclic Graphs [Flanders et al., 2011, Weisskopf et al., 2015], g-estimation methods [Moore et al., 2012], and Mendelian randomization [Relton and Davey Smith, 2015]. However, these methods omit considerations inherent to making causal statements: for example, *What randomized environmental intervention could have resulted in the data at hand?* (a conceptual stage), *What quantities are compared?* (a causal estimand definition), and *How can the non-randomized data be analyzed as emanating from a randomized intervention assignment?* (a design stage). Posing a causal question and defining the causal estimands as functions of the potential outcomes are key for objective causal inference [Rubin, 1974]. These ideas have been re-introduced to the air pollution epidemiology field by Zigler et al. [2014, 2016], who argue for well-defined actions and the use of the potential outcome framework. For years, Rubin has argued for a design stage [Rubin, 1973] and for a conceptual stage [Rubin, 2007, 2008], which was formulated more explicitly recently by Bind and Rubin [2017, 2021a], and implemented by Sommer et al. [2018, 2021, 2022].

A multistage causal inference framework

In this thesis, we present two environmental epidemiology studies whose methodologies rely on the construction of plausible hypothetical randomized experiments with the following multistage framework [Bind and Rubin, 2017]:

- *Conceptual stage:* Formulation of a causal question involving a plausible hypothetical intervention assigning an environmental exposure to units.
- *Design stage:* Reconstruction of a hypothetical randomized experiment mimicking a data structure originating from randomly assigning an environmental exposure.
- *Analysis stage:* Comparison of health outcomes, with adequate statistical methods, of units exposed to the environmental intervention to not exposed units.
- *Conclusion stage:* Interpretation of the results from the analysis stage and discussion of potentially beneficial environmental interventions, as well as recommendations for future studies.

The conceptual stage

At the conceptual stage, the causal question has to be posed. Like in traditional experimental studies, a plausible assignment of an intervention to units has to be thought of because the goal is to assess whether that intervention has any health effects by comparing the health outcomes of units under the intervention to not exposed units. The hypothetical randomized intervention assignment defines the causal estimands. The more plausible the assignment is, the more reliable the conclusions will be [Bind and Rubin, 2021b]. While conceptualizing an experiment aiming at subsequent causal analyses, the stable unit treatment value assumption (SUTVA) should be verified [Imbens and Rubin, 2015]. This assumption implies that the treatment given to any unit does not influence the outcome of other units and that, each unit exposed to the intervention, is subject to the same treatment.

The design stage

The design stage should be performed without looking at the outcome data. This stage aims at the construction of a dataset which could have plausibly been the result of an unconfounded intervention assignment. Unconfoundedness implies that the outcome of a group of units under an intervention can be compared to a control group of units, because they present, on average, “similar” background covariates, i.e., the covariates are balanced [Imbens and Rubin, 2015]. Matching is a widely-used method to create groups that are balanced. Finding covariate balance is an iterative process between matching and balance diagnostic [Rubin, 2006]. After finding a satisfying covariate balance, one can analyze a balanced data subset as if it was originated from a randomized experiment with a defined assignment mechanism.

Notice that matching strategies were suggested in the 1950s [Cochran, 1953] and have been operated “by hand” in occupational health settings in the 1970s [Tolonen et al., 1975, Hernberg et al., 1976]. However, this strategy did not spread back then because powerful computing is needed for multivariate matching and extensive balance diagnostics. When statistical software became available on personalized computers, environmental epidemiologists opted for regression modelling [Bind, 2019]. The idea of matching arose again in environmental epidemiology when case-crossover studies became popular in the 1990s. A case-crossover study [Maclure, 1991] compares the hypothesized hazardous exposures (e.g., air pollution) of a unit prior to disease onset to the exposures of this unit when it was healthy. Such design contradicts the principle of classical experimental design, since the researcher first focuses on the outcome, i.e., disease onset, and then seeks for its environmental causes. However, the “gold standard” to obtain objective inference, is to assign exposures randomly to units prior to any outcome measurement. [Rubin, 2008].

The analysis stage

Researchers feel most familiar with the analysis stage because it is the closest to model-based analyses. Before any outcome data is analyzed, the statistical analysis method to

be conducted should be defined in a protocol. After a satisfying design stage, the effect of the intervention should be estimated only once. Analysis methods can be Fisherian, Neymanian, or Bayesian [Imbens and Rubin, 2015]. A Fisherian (i.e., randomization-based) perspective [Fisher, 1935, Rubin, 1980], was chosen for our studies to circumvent relying on assumptions or asymptotic arguments.

Randomization-based inference. The central part of the analysis stage of our studies is the use of randomization-based hypothesis tests with plausible assignment mechanisms and powerful test statistics. Randomization-based inference starts with constructing the null randomization distribution of a test statistic, i.e., the distribution of values of the test statistic assuming the null hypothesis of no effect of the intervention was true [Fisher, 1935, Bind and Rubin, 2020]. Then, the Fisher p-value, i.e., the proportion of test statistics (under the null) that are as large or larger than the observed test statistic, can be calculated. A small p-value indicates the results are “worth further scrutiny” because it provides evidence for the observed test statistic to be a rare event when the null hypothesis is true. [Edgeworth, 1885, Boring, 1919, Fisher, 1925, Wasserstein et al., 2019]. Bind and Rubin [2020] highlight how simple and interpretable randomization-based inference and Fisher p-value calculation can be implemented using current computers. Surprisingly, the introduction of personal computers did not lead scientists to depart from p-value approximation based on asymptotic null randomization distribution. Therefore, in our studies, we show how to use permutations of the conceptualized intervention assignment vectors, i.e., assignment vectors following the design of our hypothetical experiments, to calculate approximate Fisher p-values. Past studies explain that when the probabilities of a unit being subject to an intervention are varying, the analysis of experiments, even when hypothetical, should reflect their design [Rubin, 2007, 2008]. This is why Bind and Rubin [2017] argue that p-values computed for observational studies can only be valid if the observational study is embedded into a plausible hypothetical randomized experiment. Accordingly, we first state a plausible randomization mechanism at the conceptual stage and implement it in the design stage before calculating a p-value.

The conclusion stage

At the conclusion stage, one should discuss statistical evidence around the results of the hypothetical experiment. It is the moment to, critically, propose how adverse effects could be curtailed by introducing some hypothetical intervention such as lower air pollution levels or smoking cessation. Once causality is suspected, the next step is to acquire medical knowledge, for instance, trying to understand biological mechanisms explaining why hazardous exposures cause unwanted health outcomes. When randomization-based inference is chosen to analyze an experiment, special care should be given to the wording of the conclusions. Alternative hypotheses should never be accepted but when p-values are small one can reject sharp null hypotheses and hint at further scrutiny of the causal question [Wasserstein et al., 2019]. With the suggested framework, results interpretation is transparent: the assumed intervention assignment mechanism is formalized at the conceptual and design stages and used in practice at the analysis stage, and conclusions are restricted

to units in the balanced data subset because results variations for units with background covariates outside our sample are unknown.

2 Two studies of adverse health effects of air pollution

Air pollution epidemiology started in the 1950s after the Great Smog of London and was established as a field with the potential to discover new medical knowledge using data collected from non-randomized studies. When statistical softwares became available on individual computers of epidemiologists at universities in the 1980s, mortality was the health outcome in the spotlight of air pollution epidemiology [Selvin et al., 1984]. Since then, many other associations have been discovered between the exposure to air pollutants and health outcomes including pulmonary and cardiovascular effects, blood markers (e.g., inflammatory, coagulation), prenatal outcomes (e.g., birth weight), and neurotoxic effects [Rückerl et al., 2011]. Some air pollution-health associations and their underlying mechanisms (e.g., for cardiovascular disease) have been extensively investigated [Brook et al., 2004]. However, other associations between environmental exposures and health outcomes provide less consensus (e.g., autism spectrum disorder, multiple sclerosis) or are simply not as much researched (e.g., gut microbiome). In this thesis, we focus on a neurological health outcome: multiple sclerosis relapses, and a newly suspected intestinal outcome: the gut microbiome [Peters et al., 2021]; two fields where environmental epidemiology research is ongoing and for which causal inference methods could propel advances.

Contrasting findings are reported on the relationship between air pollution and multiple sclerosis. Several epidemiological studies report significant associations [Oikonen et al., 2003, Gregory et al., 2008, Heydarpour et al., 2014, Angelici et al., 2016, Roux et al., 2017, Bergamaschi et al., 2017, Jeanjean et al., 2018], whereas others report no association [Palacios et al., 2017, Chen et al., 2017]. Therefore, in our “multiple sclerosis study” [Sommer et al., 2021], we use an Alsacian study population to test the hypothesis of effects of a hypothetical environmental intervention on multiple sclerosis relapses and evaluate the causal question: *Does a reduction in short-term PM_{10} levels cause a decrease in relapse occurrence risk for multiple sclerosis patients?*

Furthermore, experimental and epidemiological studies suggest that air pollution may have an effect on the gut microbiome [Mutlu et al., 2011, Kish et al., 2013, Li et al., 2015, Mutlu et al., 2018, Wang et al., 2018, Alderete et al., 2018, Liu et al., 2019, Bailey et al., 2020]. However, the lack of randomization in observational data and the complex statistical properties of microbiome data make it challenging to make causal conclusions on the associations between environment and microbiome. Therefore, in our “gut microbiome study” [Sommer et al., 2022], we use the German KORA cohort study to introduce a causal inference framework that can help investigate environment-host microbiome relationships and evaluate the causal question: *Does reducing inhaled environmental exposures alter the human gut microbiome?*

Conceptual. In both studies we conceptualized hypothetical experiments designed to study the effects of hypothetically randomized environmental interventions. Both studies

analyze the effects of an air pollution reduction intervention (e.g., randomly banning cars to be on the road during a few days a week to keep the average air pollution level below a threshold). Additionally, the gut microbiome study assesses the effects of a smoking prevention hypothetical experiment because, when adverse effects of smoking on a health outcome are established, they can support the plausibility of an adverse effect of air pollution.

Design. To limit confounding, both studies match units under the intervention to not exposed units, with respect to background covariates selected based on subject matter knowledge. We used visual and numerical balance diagnostics to compare the distribution of units pre- and post-matching.

Analysis. In the multiple sclerosis study, we do not reject the sharp null hypothesis of no effect of an air pollution reduction intervention in the overall study population. Nonetheless, in the subgroup of female patients with a Relapsing-Remitting multiple sclerosis form, a small p-value indicates that the observed intervention effect could be worth further scrutiny. In the gut microbiome study, we emphasize the importance of choosing powerful state-of-the-art test statistics for complex microbial genomic data. In air pollution reduction and smoking prevention hypothetical experiments, we reject the sharp null hypotheses of no richness, α -diversity, high-dimensional mean differences, and differential abundance for selected genera.

Conclusion. The results of our studies demonstrate that a causal inference framework can detect, with observational data, novel hypotheses on the links between environmental exposures and health outcomes. These new hypotheses are a good starting point for potential novel environmental epidemiology discoveries.

3 Outlook

A central component of this thesis is the use of a Fisherian inference approach, which we value to be a good first step to analyze untapped research questions. Our studies can be considered as stepping stones to draw causal inferences in their respective fields because we found hypotheses of environmental influences that are worth more scrutiny. The secondary analyses of the multiple sclerosis study indicate that air pollution effects on women with Relapsing-Remitting multiple sclerosis should be further studied; such subgroup of multiple sclerosis patients has, to the best of our knowledge, not yet been analyzed by environmental epidemiologists. It is known that men and women can react differently to air pollution exposures [Clougherty, 2010, Oiamo and Luginaah, 2013] but to further scrutinize our new hypothesis, a study has to be designed for this subgroup of women.

The differential abundance analyses of the smoking prevention experiment in the gut microbiome study retains a subset of genera to be further scrutinized. These genera correlate with lipid metabolites, such as serum triglycerides and high-density lipoprotein, in the same direction as previously found by Vojinovic et al. [2019]. Therefore, further studies could be done on the pathways of smoking affecting the gut and the connection with circulating metabolites and metabolic syndrome, both of which also have already been linked

to smoking [Sun et al., 2012].

Most environmental epidemiology studies provide exposure-response curves in their analysis. We chose instead to test the effects of single interventions to benefit from the appealing properties of randomized experiments. Thereby, we can directly think in terms of plausible public health interventions and potentially make applicable recommendations. Also, until now, the air pollution-multiple sclerosis or air pollution-microbiome relationships have not yet been extensively investigated (even less with causal inference methods), so a first step is to assess, whether air pollution has any effect on the health outcomes of our studies. If so, the next step could be the estimation of a causal dose response. This could be done by balancing covariates along different doses of the exposure, such as suggested in Wu et al. [2020].

At the design stage, we can only account for observed background covariates but it is possible that the assignment mechanism depends on unobserved covariates. This means that there could still be imbalances in unobserved covariates. Therefore, sensitivity analyses of how the p-value would change, had the intervention assignment been plausibly different, should be considered [Rosenbaum, 2010, Bind and Rubin, 2020]. However, for that, subject-matter knowledge on the reason why “sensitivity” p-values could deviate from the p-value calculated is needed. This idea provides material for sensitivity analyses after implementing the framework we present in our studies.

At the analysis stages of our studies, to construct the null randomization distribution and compute the p-value, we conserve the matched-pair design of the intervention assignment vector. This strategy enables avoiding making assumptions on the the underlying distribution of the data [Rubin, 1998, Bind and Rubin, 2017]. Sommer et al. [2022] call for the development of user-friendly software functions (e.g., built in R) to perform randomization-based inference while conserving the design of the intervention assignment. The use of such function would permit accountability of the design stage during the analysis of observational studies.

References

- T. L. Alderete, R. B. Jones, Z. Chen, J. S. Kim, R. Habre, F. Lurmann, F. D. Gilliland, and M. I. Goran. Exposure to traffic-related air pollution and the composition of the gut microbiota in overweight and obese adolescents. *Environmental Research*, 161:472–478, 2018.
- L. Angelici, M. Piola, T. Cavalleri, G. Randi, F. Cortini, R. Bergamaschi, A. A. Baccarelli, P. A. Bertazzi, A. C. Pesatori, and V. Bollati. Effects of particulate matter exposure on multiple sclerosis hospital admission in Lombardy region, Italy. *Environmental Research*, 145(Supplement C):68 – 73, 2016.
- M. J. Bailey, N. N. Naik, L. E. Wild, W. B. Patterson, and T. L. Alderete. Exposure to air pollutants and the gut microbiota: a potential link between exposure, obesity, and type 2 diabetes. *Gut Microbes*, 11(5):1188–1202, 2020.
- R. Bergamaschi, A. Cortese, A. Pichiecchio, F. G. Berzolari, P. Borrelli, G. Mallucci, V. Bollati, A. Romani, G. Nosari, S. Villa, and C. Montomoli. Air pollution is associated to the multiple sclerosis inflammatory activity as measured by brain MRI. *Multiple sclerosis*, 2017.
- M.-A. Bind. Causal modeling in environmental health. *Annual Review of Public Health*, 40(1):23–43, 2019.
- M.-A. Bind and D. B. Rubin. Bridging observational studies and randomized experiments by embedding the former in the latter. *Statistical Methods in Medical Research*, 28(7): 1958–1978, 2017.
- M.-A. C. Bind and D. Rubin. The importance of having a conceptual stage when reporting non-randomized studies. *Biostatistics & Epidemiology*, 5(1):9–18, 2021a.
- M.-A. C. Bind and D. Rubin. The importance of having a conceptual stage when reporting non-randomized studies. *Biostatistics & Epidemiology*, 5(1):9–18, 2021b.
- M.-A. C. Bind and D. B. Rubin. When possible, report a Fisher-exact p value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences*, 117(32):19151–19158, 2020.

- E. G. Boring. Mathematical vs. scientific significance. *Psychological Bulletin*, pages 335–338, 1919.
- R. D. Brook, B. Franklin, W. Cascio, Y. Hong, G. Howard, M. Lipsett, R. Luepker, M. Mittleman, J. Samet, S. C. Smith, and I. Tager. Air pollution and cardiovascular disease. *Circulation*, 109(21):2655–2671, 2004.
- H. Chen, J. C. Kwong, R. Copes, K. Tu, P. J. Villeneuve, A. van Donkelaar, P. Hystad, R. V. Martin, B. J. Murray, B. Jessiman, A. S. Wilton, A. Kopp, and R. T. Burnett. Living near major roads and the incidence of dementia, parkinson’s disease, and multiple sclerosis: a population-based cohort study. *The Lancet (British edition)*, 389(10070):718–726, 2017.
- J. Clougherty. A growing role for gender analysis in air pollution epidemiology. *Environmental Health Perspectives*, 118(2):167–176, 2010. ISSN 0091-6765.
- W. G. Cochran. Matching in analytical studies. *Am J Public Health Nations Health*, 43(6):684–691, 1953.
- W. G. Cochran and D. B. Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 35(4):417–446, 1973.
- F. Y. Edgeworth. Methods of statistics. *Journal of the Statistical Society of London*, pages 181–217, 1885.
- R. A. Fisher. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.
- R. A. Fisher. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.
- W. D. Flanders, M. Klein, L. A. Darrow, M. J. Strickland, S. E. Sarnat, J. A. Sarnat, L. A. Waller, A. Winquist, and P. E. Tolbert. A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*, 22(1):59–67, 2011.
- A. C. Gregory, D. G. Shendell, I. S. Okosun, and K. E. Giesecker. Multiple sclerosis disease distribution and potential impact of environmental air pollutants in Georgia. *Science of The Total Environment*, 396(1):42–51, 2008.
- S. Hernberg, M. Tolonen, and M. Nurminen. Eight-year follow-up of viscose rayon workers exposed to carbon disulfide. *Scand J Work Environ Health*, 2(1):27–30, 1976.
- P. Heydarpour, H. Amini, S. Khoshkish, H. Seidkhani, M. A. Sahraian, and M. Yunesian. Potential impact of air pollution on multiple sclerosis in Tehran, Iran. *Neuroepidemiology*, 43:233–238, 2014.
- P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

- G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA, 2015.
- M. Jeanjean, M.-A. Bind, J. Roux, J.-C. Ongagna, J. de Seze, D. Bard, and E. Leray. Ozone, NO₂ and PM₁₀ are associated with the occurrence of multiple sclerosis relapses. Evidence from seasonal multi-pollutant analyses. *Environmental Research*, 163:43–52, 2018.
- L. Kish, N. Hotte, G. G. Kaplan, R. Vincent, R. Tso, M. Gänzle, K. P. Rioux, A. Thiesen, H. W. Barkema, E. Wine, and K. L. Madsen. Environmental particulate matter induces murine intestinal inflammatory responses and alters the gut microbiome. *PLoS One*, 8(4):1–15, 2013.
- R. Li, K. Navab, G. Hough, N. Daher, M. Zhang, D. Mittelstein, K. Lee, P. Pakbin, A. Saffari, M. Bhetraratana, D. Sulaiman, T. Beebe, L. Wu, N. Jen, E. Wine, C. Tseng, J. Araujo, A. Fogelman, C. Sioutas, M. Navab, and T. Hsiai. Effect of exposure to atmospheric ultrafine particles on production of free fatty acids and lipid metabolites in the mouse small intestine. *Environ. Health Perspectives*, 123(1):34–41, 2015.
- T. Liu, X. Chen, Y. Xu, W. Wu, W. Tang, Z. Chen, G. Ji, J. Peng, Q. Jiang, J. Xiao, X. Li, W. Zeng, X. Xu, J. Hu, Y. Guo, F. Zou, Q. Du, H. Zhou, Y. He, and W. Ma. Gut microbiota partially mediates the effects of fine particulate matter on type 2 diabetes: Evidence from a population-based epidemiological study. *Environment International*, 130, 2019.
- M. Maclure. The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, 133(2), 1991.
- K. L. Moore, R. Neugebauer, M. J. van der Laan, and I. B. Tager. Causal inference in epidemiological studies with strong confounding. *Statistics in Medicine*, 31(13):1380–1404, 2012.
- E. A. Mutlu, P. A. Engen, S. Soberanes, D. Urich, C. B. Forsyth, R. Nigdelioglu, S. E. Chiarella, K. A. Radigan, A. Gonzalez, S. Jakate, A. Keshavarzian, G. S. Budinger, and G. M. Mutlu. Particulate matter air pollution causes oxidant-mediated increase in gut permeability in mice. *Particle and Fibre Technology*, 8:19, 2011.
- E. A. Mutlu, I. Y. Comba, T. Cho, P. A. Engen, C. Yazıcı, S. Soberanes, R. B. Hamanaka, R. Niğdelioğlu, A. Y. Meliton, A. J. Ghio, G. S. Budinger, and G. M. Mutlu. Inhalational exposure to particulate matter air pollution alters the composition of the gut microbiome. *Environmental Pollution*, 240:817–830, 2018.
- T. Oiamo and I. Luginaah. Extricating sex and gender in air pollution research: A community-based study on cardinal symptoms of exposure. *International Journal Of Environmental Research And Public Health*, 10(9):3801–3817, 2013.

- M. Oikonen, M. Laaksonen, P. Laippala, O. Oksaranta, E. M. Lilius, S. Lindgren, A. Rantio-Lehtimäki, A. Anttinen, K. Koski, and J. P. Erälinna. Ambient air quality and occurrence of multiple sclerosis relapse. *Neuroepidemiology*, 22:95–99, 2003.
- N. Palacios, K. Munger, K. Fitzgerald, J. Hart, T. Chitnis, A. Ascherio, and F. Laden. Exposure to particulate matter air pollution and risk of multiple sclerosis in two large cohorts of us nurses. *Environment International*, 109:64–72, 2017.
- A. Peters, T. S. Nawrot, and A. A. Baccarelli. Hallmarks of environmental insults. *Cell*, 184(6):1455–1468, 2021.
- C. L. Relton and G. Davey Smith. Mendelian randomization: applications and limitations in epigenetic studies. *Epigenomics*, 7(8):1239–1243, 2015.
- P. R. Rosenbaum. *Design of Observational Studies*. Springer, New-York, 2010.
- J. Roux, D. Bard, E. L. Pabic, C. Segala, J. Reis, J.-C. Ongagna, J. de Seze, and E. Leray. Air pollution by particulate matter (PM10) may trigger multiple sclerosis relapses. *Environmental Research*, 156:404–410, 2017.
- D. B. Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1):185–203, 1973.
- D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.
- D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- D. B. Rubin. Comment. randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593, 1980.
- D. B. Rubin. More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine*, 17(3):371–385, 1998.
- D. B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, New York, NY, USA, 2006.
- D. B. Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, 2007. ISSN 0277-6715.
- D. B. Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, 2008.
- R. Ruckerl, A. Schneider, S. Breitner, J. Cyrus, and A. Peters. Health effects of particulate air pollution: A review of epidemiological evidence. *Inhalation Toxicology*, 23(10):555–592, 2011.

- S. Selvin, D. Merrill, L. Wong, and S. T. Sacks. Ecologic regression analysis and the study of the influence of air quality on mortality. *Environ Health Perspect*, 54:333–340, 1984.
- A. J. Sommer, M. Lee, and M.-A. C. Bind. Comparing apples to apples: an environmental criminology analysis of the effects of heat and rain on violent crimes in boston. *Palgrave Communications*, 4(1):1–10, 2018.
- A. J. Sommer, E. Leray, Y. Lee, and M.-A. C. Bind. Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship. *Statistics in Medicine*, 40(6):1321–1335, 2021.
- A. J. Sommer, A. Peters, M. Rommel, J. Cyrus, H. Grallert, D. Haller, C. L. Müller, and M.-A. C. Bind. A randomization-based causal inference framework for uncovering environmental exposure effects on human gut microbiota. *PLOS Computational Biology*, 2022.
- K. Sun, J. Liu, and G. Ning. Active smoking and risk of metabolic syndrome: A meta-analysis of prospective studies. *PLoS ONE*, 7(10), 2012.
- M. Tolonen, S. Hernberg, M. Nurminen, and K. Tiitola. A follow-up study of coronary heart disease in viscose rayon workers exposed to carbon disulphide. *Br J Ind Med*, 32(1):1–10, 1975.
- D. Vojinovic, D. Radjabzadeh, A. Kurilshikov, N. Amin, C. Wijmenga, L. Franke, M. Ikram, A. Uitterlinden, A. Zhernakova, J. Fu, R. Kraaij, and C. van Duijn. Relationship between gut microbiota and circulating metabolites in population-based cohorts. *Nature Communications*, 10:Article: 5813, 2019.
- W. Wang, J. Zhou, M. Chen, X. Huang, X. Xie, W. Li, Q. Cao, H. Kan, Y. Xu, and Z. Ying. Exposure to concentrated ambient pm2.5 alters the composition of gut microbiota in a murine model. *Particle and Fibre Toxicology*, 15(1):1–13, 2018.
- R. L. Wasserstein, A. L. Schirm, and N. A. Lazar. Moving to a world beyond “ $p < 0.05$ ”. *The American Statistician*, 73(sup1):1–19, 2019.
- M. G. Weisskopf, M. A. Kioumourtzoglou, and A. L. Roberts. Air Pollution and Autism Spectrum Disorders: Causal or Confounded? *Curr Environ Health Rep*, 2(4):430–439, 2015.
- X. Wu, D. Braun, J. Schwartz, M. A. Kioumourtzoglou, and F. Dominici. Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. *Science advances*, 6(29), 2020.
- C. M. Zigler, C. Kim, C. Choirat, J. B. Hansen, Y. Wang, L. Hund, J. Samet, G. King, and F. Dominici. Causal Inference Methods for Estimating Long-Term Health Effects of Air Quality Regulations. *Res Rep Health Eff Inst*, 187:5–49, 2016.

Chapter 2

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

RESEARCH ARTICLE

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

 Alice J. Sommer^{1,2,3,5}  | Emmanuelle Leray⁴ | Young Lee¹ | Marie-Abèle C. Bind¹ 
¹Department of Statistics, Harvard University, Cambridge, Massachusetts

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

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

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

⁵Helmholtz Zentrum München – Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH)
Correspondence

Alice J. Sommer, Department of Statistics, Harvard University, One Oxford Street, Cambridge, MA 02138.

 Email: ajsommer@fas.harvard.edu

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

[Correction added on 29 January 2021 after first online publication: Correspondence details have been updated.]

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

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	255	38	108	60	147
	(28%)	(72%)	(26%)	(74%)	(29%)	(71%)
<i>MS form (at last info.)</i>						
Relapsing-remitting (n)	80	216	27	73	53	143
	(23%)	(61%)	(18%)	(50%)	(26%)	(69%)
Secondary progressive (n)	18	39	11	35	7	4
	(5%)	(11%)	(8%)	(24%)	(3%)	(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	97	31	88	4	9
	(10%)	(27%)	(21%)	(60%)	(2%)	(4%)
Type 2—incident cases (n)	63	158	7	20	56	138
	(18%)	(45%)	(5%)	(14%)	(27%)	(67%)

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 Rubin³ present, with a simple illustration, a multistaged causal inference pipeline aiming at revealing results that could have been obtained 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 stage¹⁰ 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 MS¹⁷⁻²³ 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 strategy²⁶ 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’s³ 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 aISacEP 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 aISacEP 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 definition¹⁴ 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

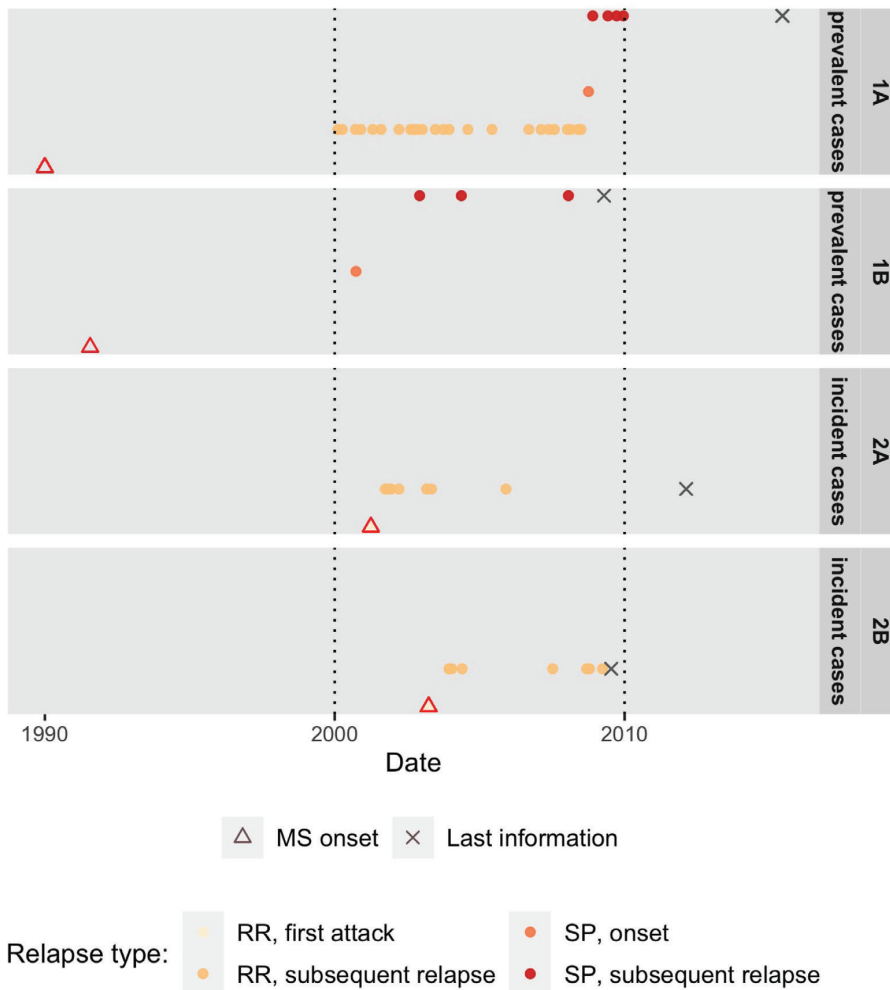


FIGURE 1 Multiple sclerosis patient types with respect to disease onset and relapse records during the study period of 2000-01-01 to 2009-10-01. Type 1—*prevalent cases*: MS onset occurred before the beginning of the study (37%); Type 1A: the patient was followed until the end of the study, Type 1B: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

period (see Table B1 in Appendix B). However, some Type 1 patients resemble Type 2 patients more (and vice versa) with 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 Wong³³ provided by the `kmeans` R function³⁴ 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.

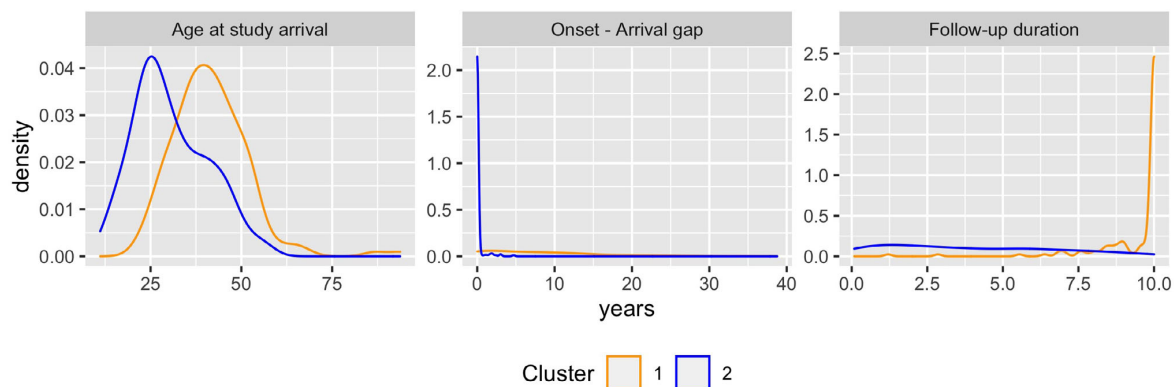


FIGURE 2 Distributions of the patient characteristics for the cluster built with K-means ($k = 2$) [Colour figure can be viewed at wileyonlinelibrary.com]

3 | METHODS

We now present the four stages of the causal pipeline³ 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 Alsatian 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 $15 \mu\text{g}/\text{m}^3$.* 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 $25 \mu\text{g}/\text{m}^3$. 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

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

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:

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

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

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

The days for which $Y_{i,s,t}^{obs} = (Y_{i,s,t}(0)|C_{i,s,t} = 1)$ can be observed are referred to as *control days*, and the days for which $Y_{i,s,t}^{obs} = (Y_{i,s,t}(1)|C_{i,s,t} = 1)$ can be observed are referred to as *treated days*.

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^* , under $W_{s,t^*}^{obs} = 0$ only if \mathbf{X}_{i,s,t^*} is “similar” to $\mathbf{X}_{i,s,t}$.

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 $\mu\text{g}/\text{m}^3$) at day $t - 6$, and $X_{i,s,t}^{(4)}$ the *maximum temperature* (in $^\circ\text{C}$) 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 $\delta_1, \dots, \delta_4$. For any pair of covariate vectors $\mathbf{X}_{i,s,t}$ and \mathbf{X}_{i,s,t^*} , we define the difference between them as

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

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 $\mu\text{g}/\text{m}^3$) are $[1, 225]$ and 63 respectively, and the range and mean of the lag-6 maximum temperature (in $^\circ\text{C}$) 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 $\delta_3 = 20 \mu\text{g}/\text{m}^3$, and the absolute difference in lag-6 maximum temperature is less than $\delta_4 = 5^\circ\text{C}$.

This constrained pair matching can be achieved by using a maximum bipartite matching³⁸ 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 $\mathbf{X}_{i,s,t}$ and \mathbf{X}_{i,s,t^*} ; and (3) an edge exists if and only if $\Delta(\mathbf{X}_{i,s,t}, \mathbf{X}_{i,s,t^*}) < +\infty$. 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 $\mathbf{X}_{i,s,t}$ 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

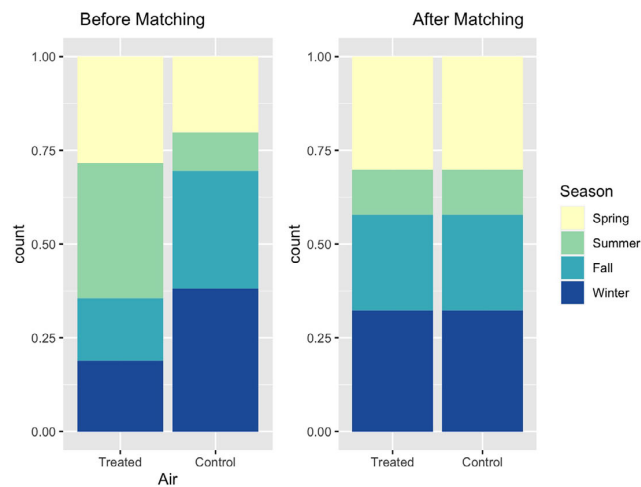
In this illustration, we follow a Fisherian analysis approach and perform hypothesis testing with a powerful test statistic comparing relapse occurrence of units subject to an intervention for air pollution reduction to units subject to higher levels of air pollution.³⁹ We do not attempt to provide an 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

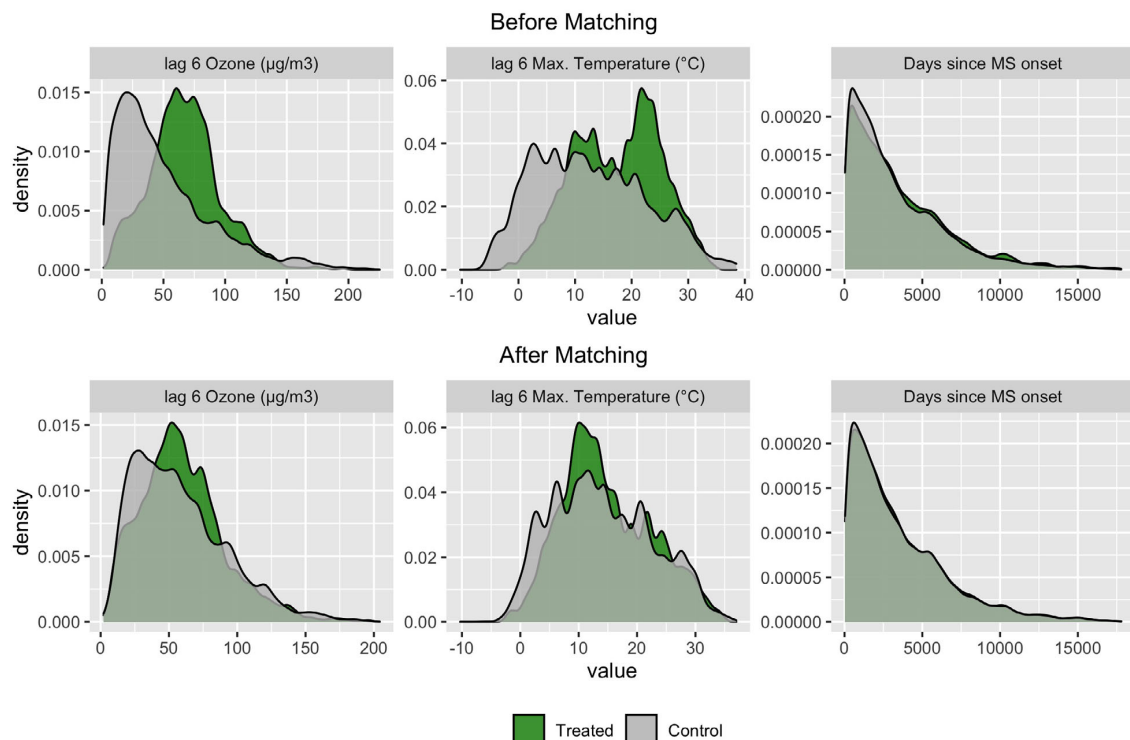
The sharp null hypothesis, stating that for each unit the intervention (exposure) has no effect on “clear days” (ie, $C_i = 1$), can be formally expressed as:

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

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



(a) Empirical distributions of the seasons among the treated and control days in the original (left panel) and the balanced subsetted (right panel) data.



(b) Empirical distributions of the pre-exposure covariates among the treated and control days in the original (upper panels) and the balanced subsetted (lower panels) data.

FIGURE 3 Empirical distributions of the preexposure covariates before and after the design stage [Colour 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 al⁴⁰ 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:

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

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 `glmex` 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: $\text{Cor}(W_{s,t}, W_{s,t-1}) = 0.5$, 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. Cautionness 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	$P\text{-value}_{CR}$	$P\text{-value}_{TC}$
	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 ($P\text{-value}_{CR} = .341$ and $P\text{-value}_{TC} = .485$, 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 ($P\text{-value}_{CR} = .038$ and $P\text{-value}_{TC} = .160$, 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's³ 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 Rosenbaum⁴⁸ 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 Rubin⁵¹ 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 al⁹ 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. Open access funding enabled and organized by ProjektDEAL.

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, Giesecker 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 history 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 Algorithm 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, Luginah 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*. 2021;40:1321-1335. <https://doi.org/10.1002/sim.8843>

APPENDIX A. DATA PREPROCESSING STEPS

1. All the patients in the aISacEP 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's¹⁴ 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 $15 \mu\text{g}/\text{m}^3$, and bigger or equal to $25 \mu\text{g}/\text{m}^3$.
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 (8)	0 (0)
Age at MS clinical onset	29 (10)	32 (11)
Age at study inclusion	39 (12)	32 (11)
Follow-up since inclusion	9 (1)	4 (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	$P\text{-value}_{CR}$	$P\text{-value}_{TC}$
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	$P\text{-value}_{CR}$	$P\text{-value}_{TC}$
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	$P\text{-value}_{B,CR}$	$P\text{-value}_{B,TC}$
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.

Supplementary Material:

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

Alice J. Sommer^{1,2,3}, Emmanuelle Leray⁴, Young Lee¹, and Marie-Abèle C. Bind^{1,*}

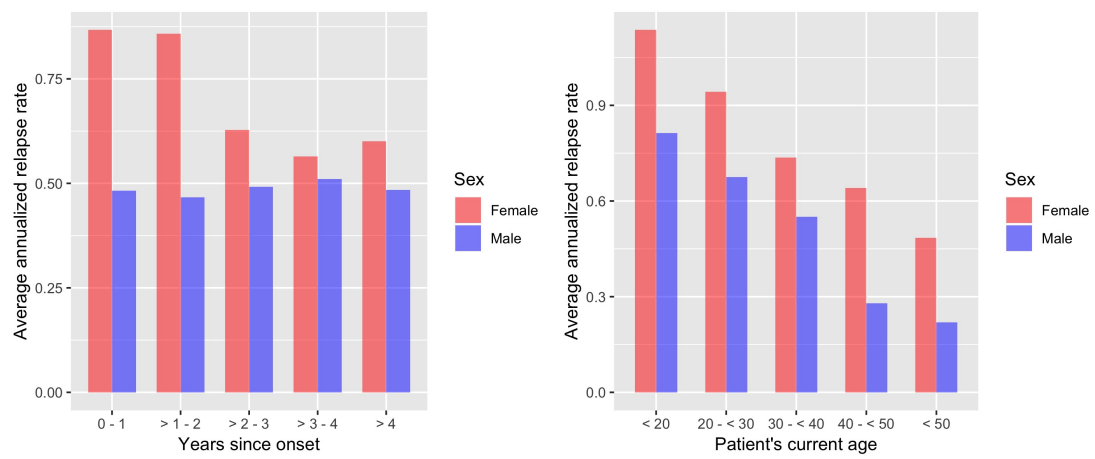
¹*Department of Statistics, Harvard University,
Faculty of Arts and Sciences, 02138 Cambridge, MA, USA*

²*Institute for Medical Information Processing, Biometry, and Epidemiology,
Faculty of Medicine, Ludwig-Maximilians-University München,
81377 Munich, Germany*

³*Institute of Epidemiology, Helmholtz Zentrum München,
85764 Neuherberg, Germany*

⁴*University of Rennes, EHESP French School of Public Health,
REPERES Pharmacoepidemiology and Health Services Research
EA, 7449, F-35000 Rennes, France*

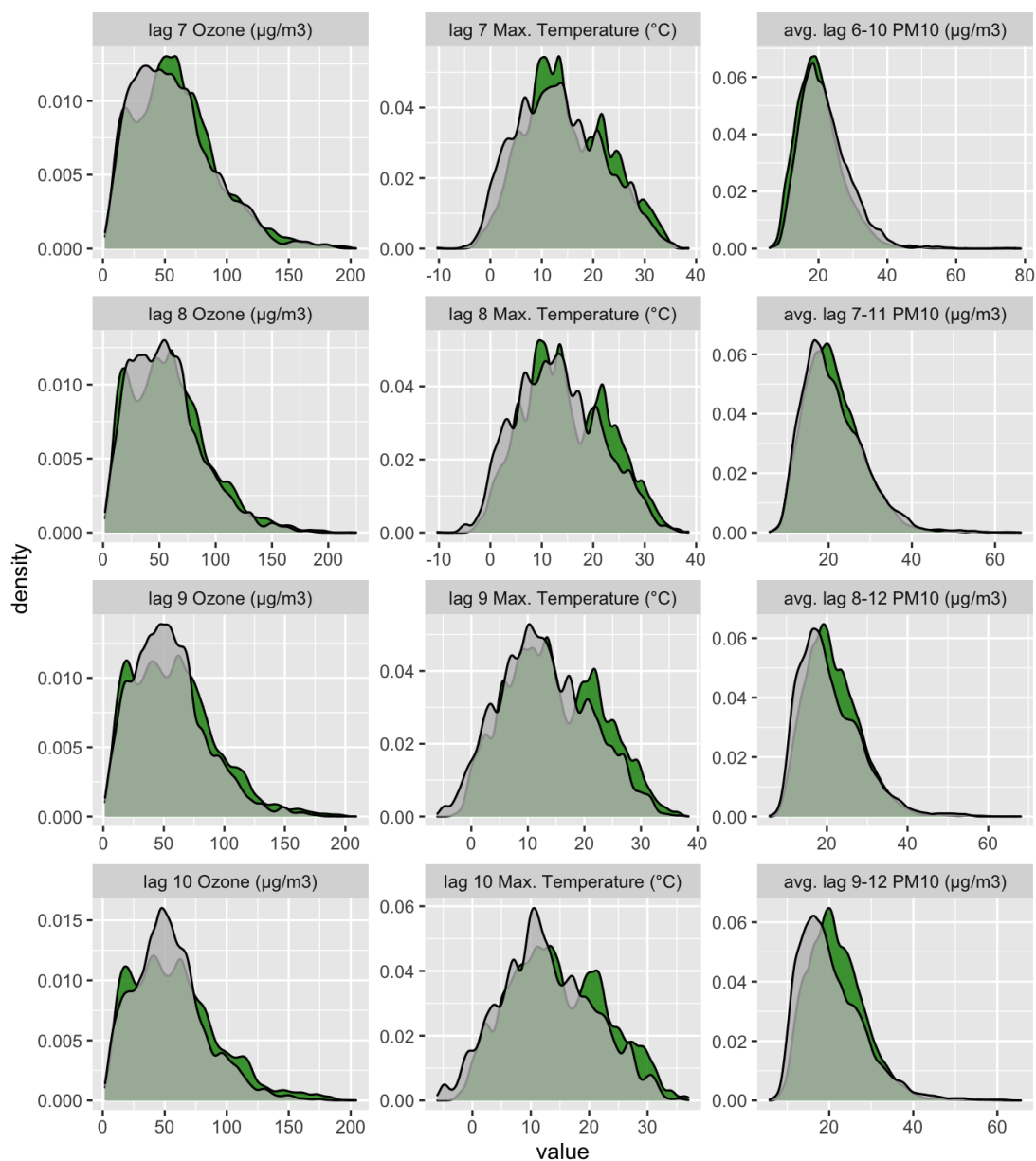
* *Corresponding author: Marie-Abèle C. Bind, ma.bind@mail.harvard.edu*



Web Figure 1: Annualized relapse rates per years at onset and patient's current age categories.

Web Table 1: Literature summary of the air pollution-MS relationship.

Authors	Outcome	Methods	Epidemiological findings
Oikinen et al. (2003) <i>Finland</i>	dichotomized monthly relapses count	multivariate logistic regression	PM associated with MS (relapse): Higher risk of MS relapse in months with highest PM_{10} (average monthly concentration in highest quartile) than lowest PM_{10} (lowest quartile).
Gregory et al. (2008) <i>Georgia</i> <i>USA</i>	MS prevalence rates	multivariate linear regression	PM associated with MS (disease onset): Higher MS prevalence rates in counties with higher long-term exposure to PM_{10} .
Heydarpour et al.(2014) <i>Teheran</i> <i>Iran</i>	case (MS patient) - control (not)	t-test	Higher long-term exposure to PM_{10} for MS cases when compared to randomly selected controls.
Angelici et al. (2016) <i>Lombardy region</i> <i>Italy</i>	hospital admission count	poisson regression	MS-related hospitalization increases on days preceded by one week with average PM_{10} levels in the highest quartile.
Bergamaschi et al. (2017) <i>Pavia province</i> <i>Italy</i>	inflammatory activity (brain MRI)	logistic regression	Higher PM_{10} levels during days before brain MRIs showing inflammatory activity in MS patients.
Jeanjean et al. (2017) <i>Alsace region</i> <i>France</i>	relapse occurrence	case-crossover study	Higher PM_{10} (O_3 , NO_2) levels during days before relapse occurrence.
Palacios et al. (2017) <i>USA</i> (Nurses Health Study I and II)	MS onset	multivariable Cox proportional hazards models	No association between average PM exposure and MS onset risk.
Hong Chen et al. (2017) <i>Ontario</i> <i>Canada</i>	MS cases	multivariable Cox proportional hazards model	No association between living near major roads and MS incidence.



Web Figure 2: Temporal unconfoundedness verification.

Chapter 3

A randomization-based causal inference framework for uncovering environmental exposure effects on human gut microbiota

A randomization-based causal inference framework for uncovering environmental exposure effects on human gut microbiota

Alice J Sommer^{1,2,3,*}, Annette Peters^{2,3,4,*}, Martina Rommel^{3,5}, Josef Cyrus³, Harald Grallert^{5,6}, Dirk Haller^{7,8}, Christian L Müller^{9,10,11,*}, and Marie-Abèle C Bind^{1,12}

¹Department of Statistics, Harvard University, Cambridge, Massachusetts, USA

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

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

⁴Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

⁵Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany

⁶German Center for Diabetes Research (DZD), München-Neuherberg, Germany

⁷ZIEL - Institute for Food & Health, Technical University of Munich, Freising, Germany

⁸Chair of Nutrition and Immunology, Technical University of Munich, Freising, Germany

⁹Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

¹⁰Department of Statistics, Ludwig-Maximilians-University München, Munich, Germany

¹¹Center for Computational Mathematics, Flatiron Institute, New York City, New York, USA

¹²Biostatistics Center, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

* alice.j.sommer@gmail.com (AJS); * peters@helmholtz-muenchen.de (AP); * cmueller@flatironinstitute.org (CLM)

May 2022

Abstract

Statistical analysis of microbial genomic data within epidemiological cohort studies holds the promise to assess the influence of environmental exposures on both the host and the host-associated microbiome. However, the observational character of prospective cohort data and the intricate characteristics of microbiome data make it challenging to discover causal associations between environment and microbiome. Here, we introduce a causal inference framework based on the Rubin Causal Model that can help scientists to investigate such environment-host microbiome relationships, to capitalize on existing, possibly powerful, test statistics, and test plausible sharp null hypotheses. Using data from the German KORA cohort study, we illustrate our framework by designing two hypothetical randomized experiments with interventions of (i) air pollution reduction and (ii) smoking prevention. We study the effects of these interventions on the human gut microbiome by testing shifts in microbial diversity, changes in individual microbial abundances, and microbial network wiring between groups of matched subjects via randomization-based inference. In the smoking prevention scenario, we identify a small interconnected group of taxa worth further scrutiny, including *Christensenellaceae* and *Ruminococcaceae* genera, that have been previously associated with blood metabolite changes. These findings demonstrate that our framework may uncover potentially causal links between environmental exposure and the gut microbiome from observational data. We anticipate the present statistical framework to be a good starting point for further discoveries on the role of the gut microbiome in environmental health.

Summary. Environmental influences on the human gut microbiome are still to be discovered or better understood. In this paper, we contribute to the field of microbiome research and environmental epidemiology by suggesting a stage-based causal inference framework relying on the foundations of the Rubin Causal Model. A particularity of the framework is the use of randomization-based inference, which we value to be a necessary exploratory inference method when tackling untapped research questions. To illustrate the framework, we explore the effects of two inhaled environmental exposures previously hypothesized to be linked with gastrointestinal diseases and the gut microbiome: air pollution exposure and cigarette smoking.

1 Introduction

The human microbiome plays a pivotal role in maintaining a healthy physiology via multiple interactions with the host. The gut microbiome, for instance, provides important metabolic capabilities for food digestion [1, 2] and regulates immune homeostasis [3]. Although dietary interventions [4], pathogen infections [5], and antibiotics use [6] can trigger rapid changes of gut microbial compositions and can lead to dysbiotic disruptions of host-microbiome interactions, the long-term impact of environmental exposures on the human gut microbiome remains poorly understood. In this paper, we provide a causal inference framework for assessing such epidemiological questions and analyze a prospective cohort with collected microbiome data. Recent technological advances, through culture-independent analyses, have facilitated a surge in observational studies of the human microbiome [7, 8, 9]. A common method to catalog microbial constituents is high-throughput amplicon sequencing [10], allowing the acquisition of gut microbiome survey data for large prospective cohort studies. Important examples include the Human Microbiome Project [11], the British TwinsUK study [12], the Dutch LifeLines-DEEP [13] and Rotterdam Studies [14], the Chinese Guangdong Gut Microbiome Project [15], the American Gut Project [16], and the German KORA study [17].

Thus far, these and other studies have linked alterations in gut microbial compositions to several common diseases, including rheumatoid arthritis, colorectal cancer, obesity, inflammatory bowel disease (IBD), and diabetes [18]. Although environmental exposures such as particulate matter (PM) [19] and smoking [20] are also related to these diseases, an understanding of environment-gut microbiome relationships and their implications for disease mechanisms has remained elusive. Here, we examine such environment-gut microbiome relationships within a causal inference framework [21] combined with state-of-the-art statistical methods for amplicon sequence variant (ASV) data [22]. We illustrate our analysis framework using data from the German KORA study [17] and focus on two inhaled environmental exposures previously hypothesized to be linked with gastrointestinal diseases and the gut microbiome: (i) particulate matter (PM) with diameter smaller or equal to 2.5 μm (PM_{2.5}) and (ii) cigarette smoking.

Air pollution exposure has been found to be associated with gastrointestinal diseases, such as appendicitis [23], inflammatory bowel disease [24], abdominal pain [25], and metabolic disorders [26]. Current research suggests that air pollution may impact the gut microbiome which, in turn, acts as a “mediator” of the association between air pollution and metabolic disorders such as obesity and type 2 diabetes [27, 28, 29]. These studies found associations between nitric oxide, nitrogen dioxide [27], PM [28], and ozone [30] exposures and the gut microbiome. Several potential pathways explain how particles affect human health. The gut is exposed to PM through: (i) mucociliary clearance, i.e., the self-cleaning mechanism of the bronchi, inducing inhaled PM to be cleared from the lungs to the gut, and (ii) oral route exposure, when food and water are contaminated by PM prior to being ingested or in the alimentary canal via inhalation [31, 32]. Results from murine studies of the effect of PM on the gut [33, 34, 35, 36, 37] suggest that exposure to PM changes the microbial composition and increases gut permeability, leading to higher systemic inflammation due to the unrestrained influx of microbial products from the gut into the systemic circulation [38].

The chemical mixture of cigarette smoke inhaled into the lungs has an effect on blood markers that, in turn, interact with the gut. Another pathway is that the toxicants of cigarette smoke swallowed into the gastrointestinal tract induce gastrointestinal microbiota dysbiosis via antimicrobial activity and regulation of the intestinal microenvironment [39]. Cigarette smoking is an inhaled exposure that has been shown to influence the susceptibility of diseases such as IBD, colorectal cancer, and systemic diseases [40, 41, 20]. Animal studies suggest that cigarette smoke may mediate its effects through alterations of intestinal microbiota [42]. In humans, shifts in the gut microbiome composition and diversity were observed after smoking cessation. These shifts were similar to previously observed shifts in obese vs. lean patients, suggesting a potential microbial link between the metabolic function of the gut and smoking cessation [43]. Comparison of the gut microbiome composition of smokers and never-smokers led to similar observations [44]. So far, the underlying mechanisms of the effect of smoking on not only gut-related, but also autoimmune diseases have not been established. It has been hypothesized that the gut microbiome may be the missing link between smoking and autoimmune diseases [20].

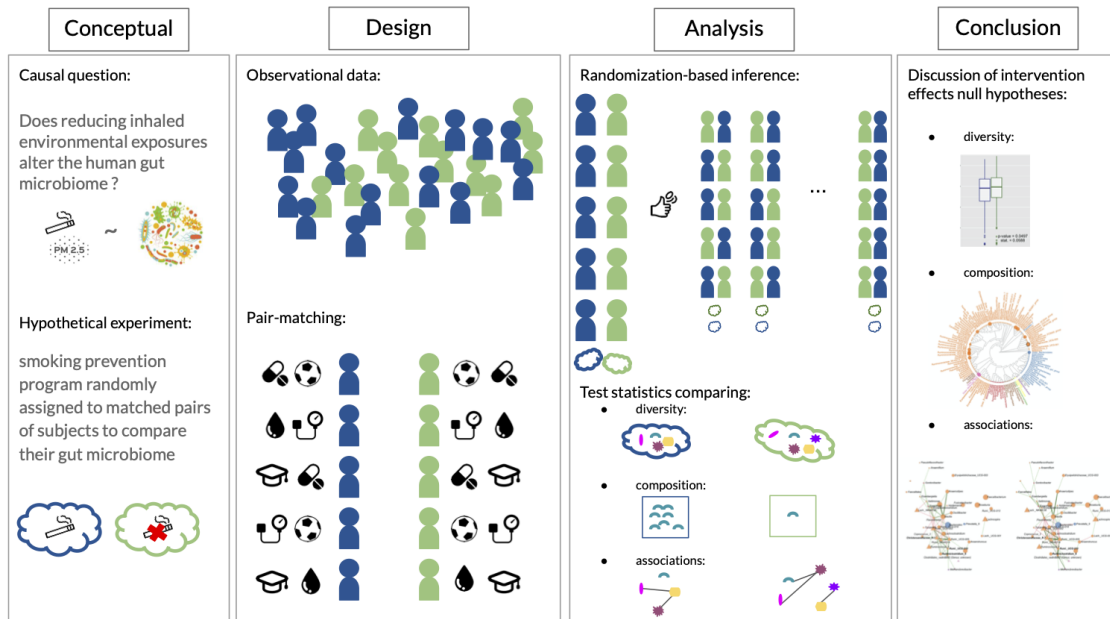


Figure 1: The four stages of the causal inference framework [21] adapted to the exploration of environment-gut microbiome relationships. Stage 1: Formulation of a plausible hypothetical intervention (e.g., decreasing inhaled environmental exposures) to examine its impacts on the gut microbiome. Stage 2: Construct a hypothetical paired-randomized experiment in which the environmental intervention been implemented randomly. Stage 3: Choose powerful test statistics comparing the gut microbiome had the subjects been hypothetically randomized to the environmental intervention vs. not and test the sharp null hypotheses of no effect of the intervention at different aggregation levels of the data. Stage 4: Interpretation of the statistical analyses and recommendations for future studies or implementation of the intervention.

Central to the present study is the investigation of the causal question: *Does reducing inhaled environmental exposures alter the human gut microbiome?* As summarized in Fig 1, we answer this question using the following four-stage analysis framework: (i) conceptualize hypothetical environmental interventions that could have resulted in the observed data at hand, (ii) design our non-randomized data, so that the unconfoundedness assumption can be assumed, (iii) choose powerful, state-of-the-art test statistics from the literature to compare human gut microbiome at different levels of taxonomic granularity between subjects assigned to the interventions vs. not, and (iv) interpret the implications of the results for recommending further studies or the studied hypothetical intervention. The reason for using this four-stage approach is for the transparency of its assumptions when interpreting results. The Methods section elaborates on each of these steps. An essential ingredient in stage (iii) of our framework is the use of a randomization-based hypothesis testing with powerful test statistics comparing subjects under an intervention vs. not [45, 46]. 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 (i.e., distribution-based) or Bayesian inferences. This Fisherian approach is a non-asymptotic first step to start shedding light on merely-touched research questions dependent on complex data structures, such as human gut microbiome data.

The present causal inference framework relies on ideas developed in the 70s [47, 48, 49, 50] and the Rubin Causal Model [51, 52] to analyze observational data by reconstructing the ideal conditions of randomized experiments, the “gold standard” to draw objective causal inferences on the effects of an intervention [53]. A formidable statistical challenge is, however, to define and test these intervention

effects for high-dimensional taxonomically-structured microbiome relative abundance data. Here, we adapted and advanced several state-of-the-art approaches from the statistical literature tailored to amplicon data, ranging from tests for α -diversity in networked communities [54, 55], Microbiome Regression-based Kernel Association Tests (MiRKAT) for β -diversity to randomization-based differential compositional mean tests [56]. We also applied and analyzed individual taxon differential abundance tests with taxonomic rank-dependent reference selection [57] and sparse compositionally robust taxon-taxon network estimation schemes [58] with novel differential edge tests [59], thus covering a comprehensive list of archetypical microbiome data analysis tasks.

Our framework complements recent causal inference approaches for microbiome data such as mediation methods [60, 61], graphical models [62], and Mendelian randomization [63, 64] to analyze observational gut microbiome data. In these studies, the target for interventions is the microbiome and the understanding of its effects on diseases, i.e., the microbiome is treated as the exposure and diseases as outcomes. Here, we are interested in examining the effects of environmental exposures (interventions) on the gut microbiome (“the” outcome), when only non-randomized data are available. To the best of our knowledge, no other observational study interested in environmental effects on the gut microbiome addressed their research question using causal inference methods.

In the following, we detail the characteristics of the KORA FF4 study population and highlight potential effects of the hypothetical interventions, air pollution reduction and smoking prevention, on the gut microbiome. In particular, we characterize potential effects in terms of changes in overall microbial diversity, taxon-level abundances, and microbial associations. In the smoking prevention analysis, we identified taxa, including *Ruminococcaceae* (*UCG-005*, *UCG-003*, *UCG-002*) and *Christensenellaceae R-7-group*, that are part of a stable sub-community in the microbial association networks and have been found to contribute to circulating blood metabolites in the LifeLines-Deep cohort [65].

2 Methods

The German KORA FF4 cohort study

The data come from the German KORA FF4 cohort study, which involves participants aged 25 to 74 years old living in the city of Augsburg [17]. The participants were subject to health questionnaires and follow-up examinations. During the study, stool samples were collected and the gut microbiota data for 2,033 participants were obtained with 16S rRNA gene sequencing. For each participant we have their long-term exposure to air pollution (particulate matter). The long-term exposure variables come from the ULTRA III study, in which air pollutants were monitored several times a year at 20 locations within the Augsburg region. From this data, annual averages of air pollutants were calculated using land-use regression models. The models explain the spatial variation of the pollutants with predictor variables derived from geographic information systems (GIS). To obtain the long-term air pollution values for each participant, land-use regression models were applied to their residential address. Moreover, to elucidate relationships between health outcomes and diet, dietary intake data were collected for 1,469 participants of the KORA FF4 cohort. Dietary intake was derived using a method combining information from a food frequency questionnaire (FFQ) and repeated 24-h food lists [66]. In brief, the usual food intake (in gram/day) was calculated as the product of the probability of consumption of a food on a given day and the average amount of a food consumed on a consumption day.

Gut microbiome data sequencing and preprocessing

DNA Extraction, 16S rRNA Gene Amplification, and Amplicon Sequencing. Fecal DNA extraction was isolated by following the protocol of [67]. The samples were profiled by high-throughput amplicon sequencing with dual-index barcoding using the Illumina MiSeq platform. Based on a study providing guidelines for selecting primer pairs [68], the V3-V4 region of the gene encoding 16S ribosomal RNA was amplified using the primers 341-forward (CCTACGGGNGGCWGCAG; bacterial domain specific) and 785-reverse (GACTACHVGGGTATCTAATCC; bacterial domain specific).

Amplification was undertaken using the Phusion High-Fidelity DNA Polymerase Hotstart as per manufacturer’s instructions. The PCR libraries were then barcoded using a dual-index system. Following a round of purification with AMPure XP beads (Beckman Coulter), libraries were quantified and pooled to 2nM. The libraries were sequenced on an Illumina MiSeq (2 x 250 bp), using facilities provided by the Ziel NGS-Core Facility of the Technical University Muenchen (TUM).

Bioinformatics. The demultiplexed, per-sample, primer-free amplicon reads were processed by the DADA2 workflow [22, 69] to infer sequence variants, remove chimeras, and assign taxonomies with the Silva v128 database [70] using the naive Bayesian classifier method [71] until the genus-level assignment and the exact matching method [72] for species-level assignment. We opted for the high-resolution DADA2 method to infer sequence variants without any fixed threshold, thereby resolving variants that differ by as little as one nucleotide. Amplicon sequence variants (ASVs) do not impose the arbitrary dissimilarity thresholds that define OTUs. They provide consistent labels because they represent a biological reality that exists outside the data being analyzed: the DNA sequence of the assayed organism, thus they remain consistent into the indefinite future [22]. The result of the DADA2 pipeline is two datasets: (i) a ASV count dataset, where each row specifies how often an ASV was sequenced and (ii) a taxonomic assignment dataset, where each row specifies the taxonomic names of an ASV. It is common to create a phylogenetic tree of the ASVs to later on calculate microbial diversity measures such as the DivNet [55] and UniFrac [73] (see the Statistical analysis stage of Methods Section 2). The multiple genome alignment for the phylogenetic tree was built with the DECIPHER R package enabling a profile-to-profile method aligns a sequence set by merging profiles along a guide tree until all the input sequences are aligned [74]. The multiple genome alignment was used to construct the *de novo* phylogenetic tree using **phangorn** R package. We first construct a neighbor-joining tree [75], and then fit a maximum likelihood tree using the neighbor-joining tree as a starting point. After 16S rRNA sequencing the 2,033 stool samples from the KORA cohort and processing the sequences with the DADA2 pipeline, we observe 15,801 ASVs (see Fig A and Table A in S1 Text).

Causal inference framework

The four stages of the causal framework [21] that we use to construct hypothetical randomized experiments to study the environment-microbiome relationship are the following:

1. *Conceptual*: Formulation of a plausible hypothetical intervention (e.g., decreasing air pollution levels) to examine its impacts on the gut microbiome.
2. *Design*: Reconstruct the hypothetical randomized experiment had the environmental intervention been implemented randomly.
3. *Analysis*: Choose valid and powerful test statistics comparing the gut microbiome had the subjects been hypothetically randomized to the environmental intervention vs. not and test the sharp null hypotheses of no effect of the intervention at different aggregation levels of the data.
4. *Summary*: Interpretation of the statistical analyses and recommendations for future studies and interventions.

Conceptual stage: formulation of the hypothetical randomized experiment in terms of potential outcomes

To understand whether environmental interventions have an effect on the human gut microbiome, the objective is to reconstruct a hypothetical experiment that mimics a controlled randomized experiment [53], in which an environmental intervention could be believed to have been randomized. Let W_i be the indicator of the assignment for subject i ($i = 1, \dots, N$) to an environmental intervention vs. none, where:

$$W_i = \begin{cases} 1 & \text{if } i \text{ is under the intervention,} \\ 0 & \text{if } i \text{ is not.} \end{cases} \quad (1)$$

The composition of a human gut microbiome can be expressed as a B -dimensional vector of the microbial abundance. We define Y_i^b as the real abundance (count) of the b^{th} bacterial taxon, $b = 1, \dots, B$ for subject i . We define the potential outcomes of subject i as $Y_i^b(1)$, the b^{th} taxon abundance (count) had subject i been randomized to the environmental intervention ($W_i = 1$), and $Y_i^b(0)$, had subject i not been randomized to the intervention ($W_i = 0$). Table 1 shows the potential outcomes for the N subjects.

<i>Taxa</i>	1		2		...	B	
<i>Subjects</i>	$W_i = 0$	$W_i = 1$	$W_i = 0$	$W_i = 1$...	$W_i = 0$	$W_i = 1$
1	$Y_1^1(0)$	$Y_1^1(1)$	$Y_1^2(0)$	$Y_1^2(1)$...	$Y_1^B(0)$	$Y_1^B(1)$
2	$Y_2^1(0)$	$Y_2^1(1)$	$Y_2^2(0)$	$Y_2^2(1)$...	$Y_2^B(0)$	$Y_2^B(1)$
...
N	$Y_N^1(0)$	$Y_N^1(1)$	$Y_N^2(0)$	$Y_N^2(1)$...	$Y_N^B(0)$	$Y_N^B(1)$

Table 1: Potential outcomes for the subjects of the hypothetical experiment

Only one of the two potential outcomes can actually be observed for each subject: this is why the Rubin Causal Model characterizes causal inference as a *missing data problem* [52], where the observed outcome of subject- i and taxa- b can be expressed as a function of both potential outcomes:

$$Y_i^{b,obs} = W_i Y_i^b(1) + (1 - W_i) Y_i^b(0) \quad (2)$$

Observed outcomes measurement

The human gut microbiome can be composed of trillions of bacteria. However, due to technology limitations, the exact abundance and number of all strains present in a human subject cannot be measured. To tackle this limitation, we opted for the processing of Amplicon Sequence Variants (ASVs) from our sequencing data to approximate the true gut microbiome composition of our study population [22, 69]. ASVs refer to individual DNA sequences recovered from a high-throughput marker gene analysis, the 16S rRNA gene in our case. Therefore, in this study the observed outcome under investigation is a $N \times A$ matrix, for $a = 1, \dots, A$ ASVs, an approximation of the $N \times B$ matrix described above. This limitation adds another layer of missing data, i.e., we are missing the true gut microbial composition of each subject. We define the ASV counts we measured for each subject- i as $C_i^{a,obs}$, which corresponds to $Y_i^{b \in A, obs}$ plus some measurement error.

Design stage: reconstruction of the conceptualized hypothetical experiment

To assess causality, randomized experiments have long been regarded as the “gold standard”. We are interested in the effect of environmental interventions that are often unpractical or ethical to assign randomly to humans within an experiment [21]. Therefore, we resort to a design stage [76] with a matched-sampling strategy to construct two hypothetical randomized experiments to assess the effects of an intervention on the changes in gut microbiome composition. The aim of our pair-matching strategy is to achieve balance in background covariates distributions as it is expected, on average, in randomized experiments. This approach attempts to create exchangeable groups as if the exposure was randomly assigned to each participant given measured covariates, to guarantee exposure assignment is not confounded by the measured background covariates. The exposure assignment mechanism determines which units receive which exposure; in other words, which potential outcomes are observed and which are missing [52]. The unconfoundedness of the assignment mechanism given covariates is a key assumption of the Rubin Causal Model.

Our pair-matching strategy aims to remove individual-specific confounding (e.g., years of age, sex, unit of BMI). Briefly, subject i under $W_i^{obs} = 1$ with pre-exposure covariates \mathbf{X}_i is matched to subject i^* , under $W_{i^*}^{obs} = 0$ only if \mathbf{X}_{i^*} is “similar” to \mathbf{X}_i . For each unit, the vector of covariates is given by $\mathbf{X}_i = (X_i^{(1)}, \dots, X_i^{(k)})$. In order to ensure covariate balance, we only allow a treated unit to be matched with a control unit if the component-wise distances between their covariate vectors are less than some pre-specified thresholds $\delta_1, \dots, \delta_k$. For any pair of covariate vectors X_i and X_{i^*} , we define the difference between them as

$$\Delta(X_i, X_{i^*}) = \begin{cases} 0 & \text{if } |X_i^{(k)} - X_{i^*}^{(k)}| < \delta_k \text{ for } k = 1, \dots, K, \\ +\infty & \text{otherwise} \end{cases}$$

This constrained pair matching can be achieved using a maximum bipartite matching [77] on a graph such that: (i) there is one node per unit, partitioned into intervention nodes and control nodes, (ii) the edges are pairs of treated and control nodes with covariates X_i and X_{i^*} , and (iii) an edge exists if and only if $\Delta(X_i, X_{i^*}) < +\infty$. By construction, using a maximum bipartite matching algorithm on this graph as implemented in the `igraph` R package produces the largest set of matched pairs that satisfy the unit-specific proximity constraints set by our thresholds. Let $N_E = \sum_{i=1}^N W_i$ be the number of subjects under the environmental intervention and $N_C = \sum_{i=1}^N 1 - W_i$ the number of control subjects, after matching.

After excluding the participants of the cohort that take antibiotics and had a cancer of the digestive organ, the pre-matched data set consists of 1,967 participants. At this stage, the objective is to create balanced data subsets for which the plausibility of the “unconfoundedness” assumption is based on a diagnostic of our choice. We choose the thresholds, $\delta_1, \dots, \delta_7$, according to the pre-matching diagnostic plots of the covariate distributions (see Figs B-G in S1 Text). We privilege a large dataset with balance, while assuring that the created pairs, or in other words “twins”, are scientifically plausible, e.g., no male and female could be matched. We assume a covariate to be balanced when its distribution is approximately the same under the exposure vs. not. The thresholds are: the absolute differences between the amount of alcohol consumption is less than $\delta_1 = 25$ g/day, between the body-mass-index is less than $\delta_2 = 4$ kg/m², between age is less than $\delta_3 = 5$ years, the diabetes status (diabetic, non-diabetic) is identical, i.e., $\delta_4 = 0$, and so are sex (male, female), i.e., $\delta_5 = 0$, and physical activity (active, inactive), i.e., $\delta_6 = 0$. Additionally, in the air pollution reduction experiment: the smoking status (smoker, ex-smoker, never-smoker) is identical, i.e., $\delta_7 = 0$, and in the smoking prevention experiment: the absolute difference between years of education is less than $\delta_7 = 3$ years.

After matching, we obtain two subsets of the data that can be analyzed as coming from two pair-randomized experiments: (i) an air pollution (ap) reduction hypothetical experiment ($N_{ap} = 198$), and (ii) a smoking prevention hypothetical experiment ($N_s = 542$); both data sets exhibit no evidence against covariate imbalance (see Table 2 and Figs B-G in S1 Text).

	Air pollution		Smoking	
	N_C	N_E	N_C	N_E
<i>Matching</i>	$PM_{2.5} \geq 13.0 \mu\text{g}/\text{m}^3$	$PM_{2.5} \leq 10.3 \mu\text{g}/\text{m}^3$	Smoker	Never smoker
Before	206	193	302	908
After	99	99	271	271

Table 2: Before and after matching number of units. The thresholds for the air pollution experiment are based on 90th and 10th percentiles of the $PM_{2.5}$ distribution.

It is well known that diet has an influence on the gut microbiome and future studies on the gut should include dietary intake data in their analysis [78, 79]. In our study, we only have access to dietary intake data for a portion of our samples, therefore we examine balance diagnostics in usual nutrient intake after matching in order to maintain a large data set before matching. Figs H-I in S1 Text show that after matching, our intervention and control units (in both hypothetical experiments) do not exhibit imbalance with respect to the following food items: potatoes/roots,

vegetables, legumes, fruits/nuts, dairy products, cereal products, meat, fish, egg products, fat, and sugar. In the same way, we checked for covariate balance after matching for medication intake, also a well-known confounder in human gut microbiome studies. Figs D and G in S1 Text show that after matching, our intervention and control units (in both hypothetical experiments) do not exhibit imbalance with respect to medication intake.

Statistical analysis stage: randomization-based inference

To compare the gut microbiome of subjects under the environmental intervention to control subjects, we choose to not rely on asymptotic arguments, but instead take a Fisherian perspective (i.e., randomization-based inference) [45, 80]. We test sharp null hypotheses (H_0) of no effect of the intervention for any unit by choosing test statistics that account for the complex microbiome data structure, including the additional “layer” of missing data. The ASV count data has a challenging structure because: (i) it is high-dimensional, (ii) some ASVs have low prevalence, (iii) the ASVs are strongly correlated, and (iv) it is compositional. ASV-count data is said to be “compositional” because between units comparison of ASV counts might not be informative due to the limited sequencing depth of the machine and the total number of sequenced reads varies from unit to unit (i.e., they have no common denominator) [81].

In randomization-based inference the goal is to construct the null randomization distribution of a test statistic assuming H_0 , T , by computing the values of the test statistic for all possible intervention assignments. Because the number of assignments is very large, we calculate an approximating p-value using N_{iter} iterations, i.e., the proportion of computed test statistics that are as large or larger than the observed test statistic: $\frac{1}{N_{iter}} \sum_{l=1}^{N_{iter}} \mathbb{1}_{T_l \geq T^{obs}}$, where $\mathbb{1}_{T_l \geq T^{obs}} = 1$ when $T_l \geq T^{obs}$, and 0 otherwise (for two-sided tests we obtain the p-values by taking absolute value of T_l and T^{obs} , i.e., $|T_l|$ and $|T^{obs}|$). A small p-value shows that the observed test statistic is a rare event when the null hypothesis is true, which indicates the results are worth further scrutiny [82]. In the following subsections, we describe the null hypotheses we test and the test statistics we use to draw randomization-based inferences with $N_{iter} = 10,000$ possible intervention assignments following a matched-pair design (see summary Table 3). This means that the permutations of the intervention assignment vectors needed to calculate the Fisher p-values follow the design of our hypothetical experiments. When units have varying probabilities of being treated, the analysis of experiments, even when hypothetical, should reflect their design [76, 53].

analysis level	data transformation	test statistic
richness	breakaway [83]	beta regression coefficient [54]
α -diversity	DivNet [55]	beta regression coefficient [54]
β -diversity	pairwise distance matrices	MiRKAT score statistic [84]
high-dimensional means	centered log ratios	mean abundance difference [56]
abundance	normalization by ratio [57]	LogFold mean difference
correlation	association matrices [58]	differential associations [59]

Table 3: Data transformation and choice of test statistics.

Diversity analyses

Within Subjects Diversity.

One of the challenges of analyzing ASV-count data is working around the low prevalence of some ASVs that are due to the limited sequencing depth of the machine and the fact that some ASVs are not shared in the entire population (see Fig A in S1 Text). Therefore, before directly testing within-subject diversity differences with so called “plug-in” estimates, it has been recently suggested to start with estimating the diversity with statistical models [54]. We will follow this idea by estimating richness with the breakaway method [83] and estimating the Shannon index for α -diversity with the DivNet method [55].

Richness. The sharp null hypothesis of no effect of the intervention on the richness can be written as: $\mathbf{H}_{0,R} : \sum_{b=1}^B \mathbb{1}_{Y_i^b(0) > 0} = \sum_{b=1}^B \mathbb{1}_{Y_i^b(1) > 0}$. To estimate the richness of subject i (i.e.,

the number of bacterial taxa present in subject i), we will estimate the total richness in subject i , observed and unobserved, by B_i with the **breakaway** model [83]. Let $f_{i,1}, f_{i,2}, \dots$ denote the number of bacterial taxa observed once, twice, and so on, in a subject i , and let $f_{i,0}$ denote the number of unobserved bacteria, so that $B_i = f_{i,0} + f_{i,1} + f_{i,2} + \dots$. The idea behind the breakaway method is that for each subject i , it predicts the number of unobserved bacteria, $f_{i,0}$, with a nonlinear regression model to, in turn, provide an estimate of B_i .

α -diversity. The sharp null hypothesis of no effect of the intervention on α -diversity can be written as: $\mathbf{H}_{0,\alpha} : \sum_{b=1}^B Y_i^b(0) = \sum_{b=1}^B Y_i^b(1)$. To have estimates for indices of the α -diversity of subject i (i.e., its total microbial abundance) and their variance, we use the **DivNet** method, because it accounts for the co-occurrence patterns (i.e., ecological networks) of bacterial taxa in the microbial community [55]. Let $Z_i^b = Y_i^b / \sum_{b=1}^B Y_i^b \in [0, 1]$ denote the unknown relative abundance of taxa b in subject i , noting that $\sum_{b=1}^B Z_i^b = 1$. As a reminder, $C_i^{a,obs}$ denotes the number of times taxa a was observed in the stool sample of subject i in our data. One of the most common α -diversity indices is the Shannon entropy [85], which is defined as: $\alpha_{i,Shannon} = - \sum_{b=1}^B Z_i^b \log(Z_i^b)$. This index captures information about both the species richness (i.e., number of species) and relative abundances of the species: as the number of species in the population increases, so does the Shannon index, and as the relative abundances diverge from a uniform distribution and become more unequal, the Shannon index decreases. In the ecological literature, researchers mostly use the following maximum likelihood estimate of $\alpha_{i,Shannon}$ (often referred to as a ‘‘plug-in’’ estimate): $-\sum_{a=1}^A \frac{C_i^a}{\sum_{a=1}^A C_i^a} \log\left(\frac{C_i^a}{\sum_{a=1}^A C_i^a}\right)$. It has been proven that this estimate is negatively biased [86]. Therefore, various corrections have been proposed and are detailed in [55]. However, most of the suggested estimates are only functions of the ASV count vectors C_i^a and do not utilize the full ASV count data matrix C and the co-occurrence pattern, i.e., ecological network, of the ASVs. Willis and Martin [55] showed that these networks can have substantial effects on estimates of diversity and proposed an approach, called **DivNet**, to estimating diversity in the presence of an ecological network. **DivNet** estimates are based on log-ratio transformations by fixing a ‘‘baseline’’ taxon for comparison, which are modeled by a multivariate normal distribution to incorporate the co-occurrence structure between the taxa as the covariance matrix. The main advantage of **DivNet** method is the use of information shared across all samples to obtain more precise and accurate estimates.

Choice of test statistic. The test statistic we use to test $\mathbf{H}_{0,R}$ and $\mathbf{H}_{0,\alpha}$ are the coefficient of the intervention indicator estimated by the regression suggested by Willis et al. [54]. Using the coefficient of a model as the test statistic of a Fisher test was introduced in the 70s [87]. 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 [47, 48].

Willis et al. [54] suggest to test changes in richness (B_i) and α -diversity ($\hat{\alpha}_i$) with a hierarchical regression model, assuming that richness is a function of: the intervention indicator W_i , random variation that is not attributed to the covariates, and the standard error previously estimated with breakaway or **DivNet** (because not every bacterial taxon in each subject was observed so we cannot not know the true richness or α -diversity for any i). The regression models are built with the **betta** function available in the **breakaway** R package [83, 54].

Between Subjects Diversity.

β -diversity. Distance-based analysis is a popular approach for evaluating the association between an exposure and microbiome diversity. The pairwise distances, d_{ii^*} , for high-dimensional data we consider are the: UniFrac (unweighted) distance [73], Jaccard index, Aitchison distance [88] (i.e, Euclidean distance on centered log-ratio transformed data), and Gower distance [89] (on centered log-ratio transformed data). We choose the unweighted paired UniFrac, because it is a distance metric (i.e., a non-negative real-valued function) as opposed to the generalized UniFrac. In the same way, the Jaccard distance was chosen as opposed to the commonly used Bray-Curtis. The sharp null hypothesis of no effect of the intervention on β -diversity can be written as: $\mathbf{H}_{0,\beta} : \mathbf{d}_{ii^*}(0) = \mathbf{d}_{ii^*}(1)$.

Choice of test statistic. Despite the popularity of distance-based approaches, the field of microbiome studies suffers from technical challenges, especially in selecting the best distance. Therefore, we use the suggested microbiome regression-based kernel association test (**MiRKAT**) [84]

that uses a kernel regression and a standard variance-component score test statistic [90]. To consider different distance measures, the optimal MiRKAT: tests $\mathbf{H}_{0,\beta}$ for each individual kernel, obtains the p-value for each of the tests, and then adjust for multiple comparison with a p-value with an omnibus test. Instead, we use a fully randomization-based multiple comparison adjustment method detailed subsequently.

Multiple comparison adjustments. We follow the fully randomization-based procedure for multiple comparisons adjustments suggested by Lee et al. [91], which is directly motivated by the intervention assignment actually used in the experiment. This procedure has been suggested to have sufficient power to detect causal effects [91]. In our hypothetical experiments, we have matched paired intervention assignments. Both the unadjusted and adjusted p-values in the procedure are randomization-based, so do not require any assumptions about the underlying distribution of the data. The *adjusted* p-values are calculated following Steps 1-4:

1. Calculate for each hypothesis h , an unadjusted p-value for the observed test statistic by taking the proportion of computed test statistics that are as large or larger than the observed test statistic. This procedure is detailed in the introduction of the Statistical analysis stage section. Also, for each hypothesis h , $h = 1, \dots, H$, and intervention assignment iteration $iter$, $iter = 1, \dots, N_{iter}$, record the vector of calculated test statistics $T_\beta^{h,iter} = (T_\beta^{1,1}, \dots, T_\beta^{H,N_{iter}})$.
2. For each h and each iteration $iter$, calculate an unadjusted randomization-based p-value, with $T_\beta^{h,iter}$ as the observed test statistic. For each $iter$, record the minimum p-value of the H p-values.
3. The repetitions of Step 2 capture the joint randomization distribution of the test statistics and thus, of the unadjusted p-values.
4. To calculate the adjusted p-values for the observed test statistics, for each h , take the proportion of “minimum p-values” (recorded in Step 2) that are less than or equal to its unadjusted p-value calculated in Step 1.

Step 2-3. essentially represent a translation of the multiple test statistics into p-values sharing a common 0-1 scale.

Composition analyses

Compositional equivalence.

The compositionality problem means that: a change in abundance (i.e., sequenced counts) of a taxon in a sample induces a change in sequenced counts across all taxa. This problem, among others, leads to many false positive discoveries when comparing taxon abundances between groups. Moreover, because the components of a composition must sum to unity, directly applying standard multivariate statistical methods intended for unconstrained data to compositional data may result in inappropriate and misleading inferences [88]. Therefore, we impose a centered log-ratio transformation of the compositions before testing the null hypothesis of no difference in average microbial abundance as suggested by [56].

For the measured microbiome data C , the centered log-ratio matrices $L = (L_1, \dots, L_N)$ are defined by $L_i^a = \log\left(\frac{C_i^a}{g(\mathbf{C}_i)}\right)$, where $g(\mathbf{C}_i) = (\prod_{a=1}^A C_i^a)^{1/A}$ denotes the geometric mean of the vector $\mathbf{C}_i = (C_i^1, \dots, C_i^A)$. The sharp null hypothesis of no microbiome composition difference between the subjects under the intervention vs. not can be written as $\mathbf{H}_{0,\mathbf{M}}$: for each subject i , $L_i(0) = L_i(1)$.

Choice of test statistic. The scale invariant test statistic suggested by [56] for testing $\mathbf{H}_{0,\mathbf{M}}$ is based on the differences $\bar{L}_E^{a,obs} - \bar{L}_C^{a,obs}$, where $\bar{L}_E^{a,obs} = 1/N_E \sum_{i:W_i=1} L_i^a$ is the sample mean of the centered log ratios for subjects under the intervention. Because microbiome data are often sparse (i.e., only a small number of taxa may have different mean abundance), the following test statistic is considered: $T_M = \frac{N_E N_C}{N_E + N_C} \max_{1 \leq a \leq A} \frac{(\bar{L}_E^{a,obs} - \bar{L}_C^{a,obs})^2}{\hat{\gamma}_{aa}}$, where $\hat{\gamma}_{aa}$ are the pooled-sample centered log-ratio variances.

Differential abundance

The compositional nature of the microbiome data requires to choose appropriate reference sets with respect to which testing of changes in individual taxon relative abundances becomes feasible [81]. A recent approach that follows this methodology is the DACOMP (differential abundance testing with **com**positionality adjustment) method, proposed by [57]. DACOMP is a data-adaptive approach that: 1) identifies a subset of non-differentially abundant (reference) ASVs (R) in a testing dataset, and 2) tests the null of no differential abundance (DA) of the other ASVs (a) “normalized-by-ratio” in a training dataset. First, a taxon enters the set $R = (r_1, \dots, r_F)$ if it has low variance (< 2) and high prevalence ($> 90\%$) (see Figs L-M in S1 Text). For the analyses at the ASV level, we chose the variance to be < 3 and the prevalence to be $> 40\%$ as thresholds in order to have at least one reference per subject. Second, using the suggested “normalization-by-ratio” approach, the null hypothesis to be tested for ASV a is that ASV a is not differentially abundant: $\mathbf{H}_{0,DA}^{(a \notin R)}$: $\frac{C_i^a(0)}{C_i^a(0) + \sum_{f=1}^R C_i^{r_f}(0)} = \frac{C_i^a(1)}{C_i^a(1) + \sum_{f=1}^R C_i^{r_f}(1)}$,

Choice of test statistic. To test this sharp null hypothesis, we use the LogFold change available in the `dacomp` package with the `Compute.resample.test` function. This function is useful to perform randomization-based inference for differential abundance testing, because it enables to directly incorporate a matrix of hypothetically randomized intervention assignments, which is an appealing feature when researchers work with particular designs. Because we are testing $\mathbf{H}_{0,DA}^{(a \notin R)}$ $\|A\| - \|R\|$ times at all taxonomic ranks, we adjust for multiple tests with the method described in the β -diversity analysis section [91].

Partial correlation structure

For our matched intervention and control subjects, we predicted microbial association networks using the Sparse Inverse Covariance estimation for Ecological ASsociation Inference (SPIEC-EASI) framework [58] that uses 1) centered log-ratio transformations of the observed ASV counts, $C_i^{a,obs}$, to perform 2) Sparse Inverse Covariance selection (with the graphical lasso method [92]), and finally 3) pick a model based on edge stability (with the StARS method [93]) to obtain a sparse inverse covariance matrix. The non-zero entries of this matrix are proportional to the negative partial correlations among the taxa and form the edge set in an undirected weighted graph $G = (V, E)$. Here, the vertex (or node) set $V = v_1, \dots, v_p$ represents the p genera and the edge set $E \subset V \times V$ the possible associations among them. The null hypotheses of no effect of the environmental intervention on the observed genera network associations can be expressed as: $\mathbf{H}_{0,N} : E(0) = E(1)$.

Choice of test statistic. We compare the intervention and control networks with test statistics for the difference in genera associations individually. To generate sampling distributions of the test statistics under $\mathbf{H}_{0,N}$, the intervention and control labels are reassigned 10,000 times to the samples while the matched pair structure is maintained, i.e., the assignment to intervention or control is permuted within each pair. The SPIEC-EASI framework is then re-applied to each permuted data set. This procedure is implemented with the Network Construction and Comparison for Microbiome Data, `NetCoMi`, R package [59]. To adjust for multiple differential association tests, we use the method described in the β -diversity and differential abundance analyses section [91].

Summary stage: interpretation of the results

If the null hypothesis of no difference in the gut microbiome 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 treatments, assuming the assignment “unconfoundness” assumption holds. We can then report that the gut microbiome composition was or was not altered by the introduction of the environmental intervention. 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 Figs B-I in S1 Text). The data do not provide direct information for “unmatched” units. Caution regarding extrapolation to units with covariate values beyond values observed in the balanced subset of the data is necessary.

3 Results

To illustrate our causal inference framework, we first conceptualize two hypothetical environmental interventions that potentially influence the gut microbiome: (i) an air pollution reduction, and (ii) a smoking prevention intervention. Second, for each intervention, we construct a hypothetical matched-pair randomized experiment, aiming at satisfying the “unconfoundedness” assumption (see Methods section). Third, we analyze the “unconfounded”/“as-if randomized” data subset with randomization-based inference to test sharp null hypotheses of no effect of the interventions for each unit at different taxonomic levels of the microbial ASV data. The results presented subsequently correspond to the third stage of the framework. Fourth, causal conclusions are developed in the Discussion section. Following the American Statistical Association statement [94, 82], we avoid searching for “statistically significant” results with a dichotomous approach. To give structure to our results reporting, we reject the sharp null hypotheses of no effect of an environmental intervention when the p-value is lower or equal to 0.1 or, when computed, when the adjusted p-value is lower or equal to 0.2. We are more tolerant with adjusted p-values because multiple comparison adjustments are conservative and our study is exploring a nearly untapped field. Nonetheless, we highly recommend to the readers interested in our research questions or result replication to examine all reported p-values in Figs and Tables, because higher p-values do not mean that an effect is improbable, absent, false, or unimportant [82].

Characteristics of study population

Our study is based on data from the KORA FF4 study cohort [17]. Because we performed a design stage before analyzing the data we have two study populations, one per hypothetical experiment, which are subsets of the entire cohort (see Design stage in the Methods section). In the air pollution reduction experiment, we analyze 99 matched pairs of subjects living in highly ($PM_{2.5} \geq 13.0 \mu\text{g}/\text{m}^3$) and less ($PM_{2.5} \leq 10.3 \mu\text{g}/\text{m}^3$) polluted areas with similar background characteristics distributions (Table 4 and Figs B-D and Fig H in S1 Text). The thresholds for the air pollution experiment intervention are based on 90th and 10th percentiles of the $PM_{2.5}$ distribution. We focus on the $PM_{2.5}$ pollutant, originating mainly from traffic emissions and fossil fuel combustion, for its known penetrating effects into the lung and potential implication for the gut microbiome [27]. In the smoking prevention experiment, we analyze 271 matched pairs of smokers and never-smokers (with background characteristics distributions presented in Table 4 and Figs E-G and Fig I in S1 Text). A total of 45 units are included in the balanced data subset of both hypothetical experiments.

		Air pollution (PM _{2.5})				Smoking			
		≥ 13.0 μg/m ³		≤ 10.3 μg/m ³		Smoker		Never-Smoker	
		Mean	St. d.	Mean	St. d.	Mean	St. d.	Mean	St. d.
Age		60.6	12.4	60.3	12.4	54.2	9.4	54.4	9.6
Body Mass Index		27.0	4.3	27.0	3.8	26.7	4.4	26.7	4.2
Alcohol intake (g/day)		11.3	14.1	11.5	13.9	13.0	15.6	11.6	14.3
Years of education		11.9	2.6	11.7	2.8	11.7	2.3	11.8	2.2
		N	%	N	%	N	%	N	%
Sex	F	41	20.7	41	20.7	130	24.0	130	24.0
	M	58	29.3	58	29.3	141	26.0	141	26.0
Smoking	Ex-S.	27	13.6	27	13.6	-	-	-	-
	Never-S.	62	31.3	62	31.3	-	-	-	-
	Smoker	10	5.1	10	5.1	-	-	-	-
Diabetes	No	95	48.0	95	48.0	264	48.7	264	48.7
	Yes	4	2.0	4	2.0	7	1.3	7	1.3
Phys. Activity	No	36	18.2	36	18.2	125	23.1	125	23.1
	Yes	63	31.8	63	31.8	146	26.9	146	26.9

Table 4: Baseline characteristics of the study population in the air pollution reduction (left table) and smoking prevention experiments (right table). Continuous variables: mean and standard deviation (St. d.). Categorical variables: number of samples per category (N) and proportion of category (%).

Microbial diversity analysis

A common first step in microbiome data analysis is estimating and assessing microbial diversity. We begin by investigating the potentially causal effects of the interventions on within-subject diversity (α -diversity) and between-subject variation (β -diversity), respectively.

Within-subject diversity

Gut bacterial richness and Shannon diversity were estimated on the ASV level with the breakaway [83] and DivNet [55] method, respectively. Comparisons of the distributions of these estimated variables between subject under the intervention vs. not in both hypothetical experiments are shown by boxplots in Fig 2. The small approximate Fisherian p-values based on 10,000 permutations of the intervention assignment give us ground for rejecting the null hypotheses of no effect of an air pollution reduction (p-value_{ap,richness} \approx 0.0008, p-value_{ap, α -div.} \approx 0.0388) and smoking prevention (p-value_{s,richness} \approx 0.1518, p-value_{s, α -div.} \approx 0.0497) on the diversity of the human gut microbiome. On average, lower diversity was observed in the subjects living in polluted areas or smokers compared to participants living in less polluted areas or non-smokers. This diversity difference motivates the more in-depth analyses of the gut microbiome composition presented subsequently.

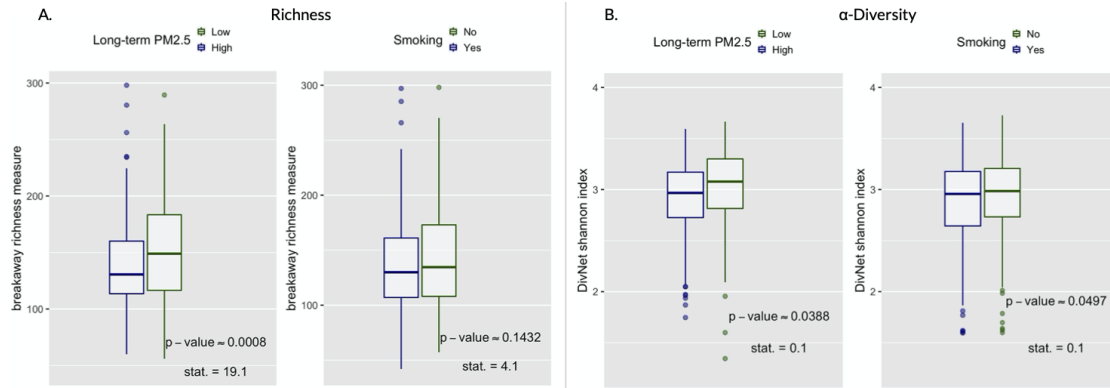


Figure 2: Richness and α -diversity. Boxplots (with median), values of the test-statistics from the **beta** regression [54], and one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design. (A) Boxplots of the richness. (B) Boxplots of the α -diversity.

Between-subject variation

To estimate β -diversity indices, we calculated UniFrac, Aitchison, Jaccard, and Gower dissimilarities between all possible pairs of subjects. The results are shown in Table 5. To alleviate the problem of choosing the best dissimilarity metric for β -diversity estimation, we follow the Microbiome Regression-based Kernel Association Test (MiRKAT) of Zhao et al. [84] suggesting to compute several metrics and then adjust for multiple comparisons. In both experiments, we reject the sharp null hypotheses of no effect of the intervention on between-subject variation.

<i>distance</i>	Air pollution			Smoking		
	test-statistic	p-value	p-value _{adj}	test-statistic	p-value	p-value _{adj}
UniFrac	12.1	0.0199	0.0506	61.5	0.0024	0.0070
Aitchison	82596.0	0.1096	0.2466	356921.5	0.0001	0.0003
Jaccard	19.4	0.0884	0.2043	84.5	0.0001	0.0003
Gower	0.2	0.0089	0.0250	0.1	0.0485	0.1204

Table 5: β -diversity. Microbiome Regression-based Kernel Association Test (MiRKAT), unadjusted and adjusted one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

Microbial compositions analysis

We next investigated whether shifts in microbial compositions as a whole or differences in specific microbial taxa were observable in the hypothetical experiments. We illustrate this by designing and analyzing sharp null hypotheses for global compositional means and differential genus abundances.

Compositional mean differences

Testing whether two study groups have the same microbiome composition can be viewed as a two-sample testing problem for high-dimensional compositional mean equivalence. We tested sharp null hypotheses using a test statistic developed particularly for that purpose by Cao et al. [56]. Table 6 summarizes the results for each taxonomic level. We reject the sharp null hypotheses of gut microbiome composition equivalence for the air pollution reduction and smoking prevention experiments. In both experiments, p-values are higher at the ASV level than at higher taxonomy levels.

		ASV	Species	Genus	Family	Order	Class	Phylum
Air Pollution	nb. of taxa (p)	4,370	414	252	74	44	29	15
	test statistic	12.8	12.9	11.9	8.8	8.4	8.4	8.1
	p-value	0.1451	0.0722	0.0733	0.1521	0.1161	0.1021	0.0591
Smoking	nb. of taxa (p)	7,409	479	271	81	48	31	16
	test statistic	13.0	14.5	13.3	11.6	8.6	9.4	10.4
	p-value	0.1607	0.0302	0.0384	0.0279	0.0859	0.0440	0.0135

Table 6: Compositional equivalence test. Test statistic for high-dimensional data suggested by [56] and one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

Differential taxon abundances

For compositional microbiome data, identifying sets of potentially “differentially abundant taxa” relates to testing sharp null hypotheses of no difference in abundance of individual taxa with respect to a reference set. We conducted such an analysis on the genus level for all genera present in at least 5% of the samples. This prevalence threshold was guided by the amount of information preserved when performing filtering, i.e., microbial abundance and the number of taxa observed per sample (see Figs N-Q in S1 Text). We applied the Differential abundance testing for compositional data (DACOMP) approach [57] and used two-sided tests since we lack prior knowledge on the direction of the abundance changes. Fig 3 highlights the key DACOMP results for both experiments. In the air pollution reduction experiment, we reject the sharp null hypothesis of no differential abundance only for the *Marvinbryantia* genus ($p\text{-value}_{adj.} = 0.0120$) (see Table B in S1 Text). We also reject the sharp null hypothesis of no effect of smoking prevention for eleven genera (see Fig 3 and Table C in S1 Text). Five belong to the *Ruminococcaceae* family: *Ruminococcaceae-UCG-002*, *Ruminococcaceae-UCG-003*, *Ruminococcaceae-UCG-005*, *Ruminococcus-1*, and *Ruminococcaceae-NK4A214-group*, three to the *Lachnospiraceae* family: *Lachnospira*, *Lachnospiraceae-NK4A136-group*, and *Coprococcus-1*, one to the *Christensenellaceae* family: *Christensenellaceae-R-7-group*, and two to the *Mollicutes* class, which belong to the *NB1-n* and *Mollicutes-RF9* order.

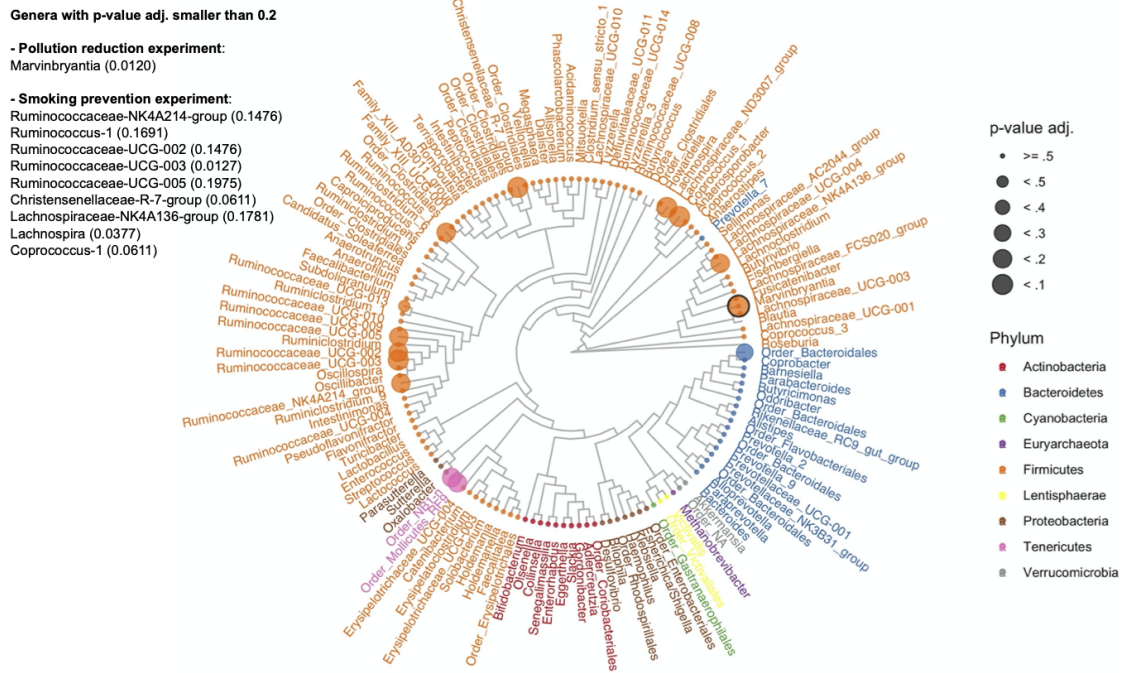


Figure 3: Differential abundance. For each genus, adjusted two-sided randomization-based p-values for 10,000 permutations of the smoking prevention intervention assignment following a matched-pair design. Genera with no tip point belong to the set of reference taxa. Black circled tip point: differentially abundant genus (*Marvinbryantia*) in the air pollution reduction experiment.

Microbial network analysis

To gain insights into changes in the organizational structure of the underlying microbial gut ecosystem, we next calculated sparse genus-genus association networks for each exposure level and hypothetical experiment and highlight the results of our randomization-based differential association testing.

Genus-genus association networks

We used the Sparse Inverse Covariance estimation for Ecological Association Inference (SPIEC-EASI) framework [58] to infer genus-genus associations in our two hypothetical experiments. We used the glasso mode of SPIEC-EASI with default parameters (see Methods for details). Fig 4A shows the overall structure of the learned sparse association networks for the smoking prevention experiment (smokers (left panel) and non-smokers (right panel), respectively). Each network possesses a single large connected component consisting of 30-40 mostly *Firmicutes* genera (highlighted area in Fig 4A). These connected components also included the majority of the previously identified potentially differentially abundant genera, including *Ruminococcaceae* (*UCG-005*, *UCG-002*), *Ruminococcus-1*, and *Christensenellaceae-R-7-group* (see Fig 4B for a detailed view of the connectivity pattern). The genus-genus associations networks derived from the air pollution reduction experiment showed similar overall topological features containing one large connected component of 60 genera, including *Ruminococcaceae* (*UCG-005*, *UCG-003*, *UCG-002*) and *Christensenellaceae-R-7-group* among others (see also Fig R in S1 Text).

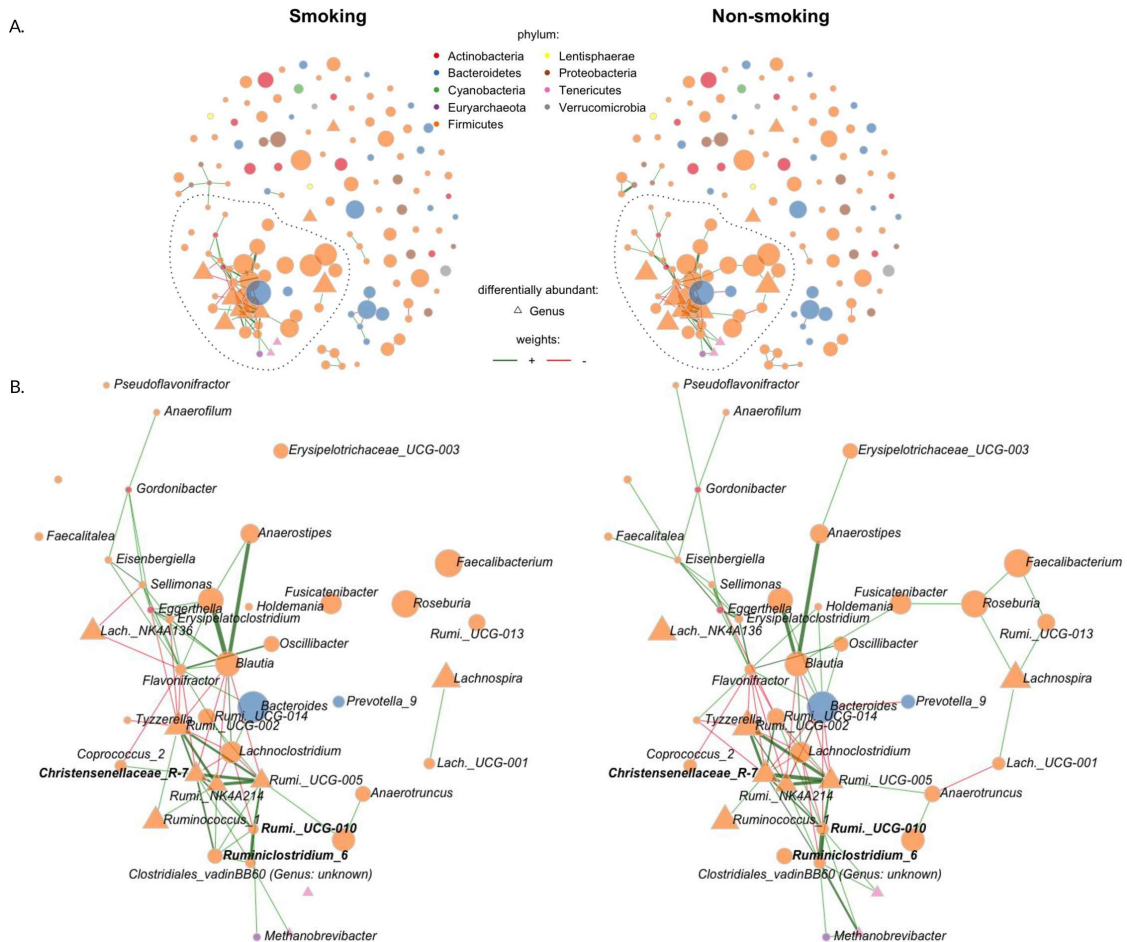


Figure 4: Genus-genus associations of smokers and never-smokers ($n = 271$, $p = 140$). (A) Visualization of the genus-genus partial correlations estimated with the SPIEC-EASI method. Edges thickness is proportional to partial correlation, and color to sign: red: negative partial correlation, green: positive partial correlation. Node size is proportional to the centered log ratio of the genus abundances, and color is according to phyla. Triangle shaped nodes are differentially abundant (see Figure 3). (B) Zoom in largest connected component and differential associations (bold genera).

Differential genus-genus associations

To identify potentially differential network associations in the intervention experiments, we coupled the SPIEC-EASI network estimation procedure with permutations of the intervention assignment, available in the NetCoMi R package [59] (see also Methods for details). For each hypothetical experiment, we list the five genus-genus associations with smallest adjusted two-sided randomization-based p-values in Table 7 and highlight these associations in Fig 4B. In the air pollution reduction experiment, we reject the sharp null hypothesis of no differential association for two edges: the *Succinivibrio/Slackia* edge ($p\text{-value}_{adj.} \approx 0.0661$), and the *Ruminiclostridium/Cloacibacillus* edge ($p\text{-value}_{adj.} \approx 0.1063$) (see Table 7 and Fig R in S1 Text).

Air pollution	
Genus-genus associations (- : disappearance after intervention)	p-value _{adj}
<i>Succinivibrio/Slackia</i> (-)	0.0661
<i>Ruminiclostridium/Cloacibacillus</i> (-)	0.1063
<i>Cloacibacillus/Lachnospiraceae-FCS020-group</i>	0.2795
<i>Megasphaera/Alistipes</i>	0.4147
<i>Bacteroidales</i> (Genus: unknown)/ <i>Prevotella-2</i>	0.4753
Smoking	
Genus-genus associations (- : disappearance after intervention)	p-value _{adj}
<i>Christensenellaceae-R-7/Ruminiclostridium-6</i> (-)	0.1585
<i>Ruminococcaceae-UCG-010/Ruminiclostridium-6</i> (-)	0.1585
<i>Ruminococcaceae-UCG-014/Flavonifractor</i>	0.2031
<i>Clostridiales-vadinBB60/Ruminiclostridium-6</i>	0.2376
<i>Ruminococcaceae-UCG-013/Faecalibacterium</i>	0.2492

Table 7: Differential associations of genera. Smallest five adjusted two-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

In the smoking prevention experiment, we also reject the sharp null hypothesis of no differential association for two edges: the *Ruminiclostridium-6/Ruminococcaceae-UCG-010* edge (p-value_{adj.} ≈ 0.1585), and the *Ruminiclostridium-6/Christensenellaceae-R-7-group* edge (p-value_{adj.} ≈ 0.1585) (see Table 7). The genera that participate in these potentially differential associations are also highlighted in Fig 4B.

Exploring associations between genera and lipid metabolites

The gut microbiome is a substantial driver of circulating lipid levels, and prior work has shown [95, 96, 65] that the relative abundance of several microbial families, including *Christensenellaceae*, *Ruminococcaceae*, and the Tenericutes phylum were negatively correlated with triglyceride and positively associated with high-density lipoproteins (HDL) cholesterol. Since our analysis identified a small interconnected group of genera, including *Christensenellaceae* and *Ruminococcaceae*, for whom we rejected the no differential abundance hypothesis, we performed an exploratory data analysis to investigate taxa-serum lipid measurements associations. Four lipids were measured in blood serum samples of our study population from the KORA cohort: total, HDL, and LDL, cholesterol, as well as triglyceride levels. Fig 5A shows the correlation between these lipids and the genera we discovered in our hypothetical experiments. Tendencies similar to those reported in previous studies can be observed in our data.

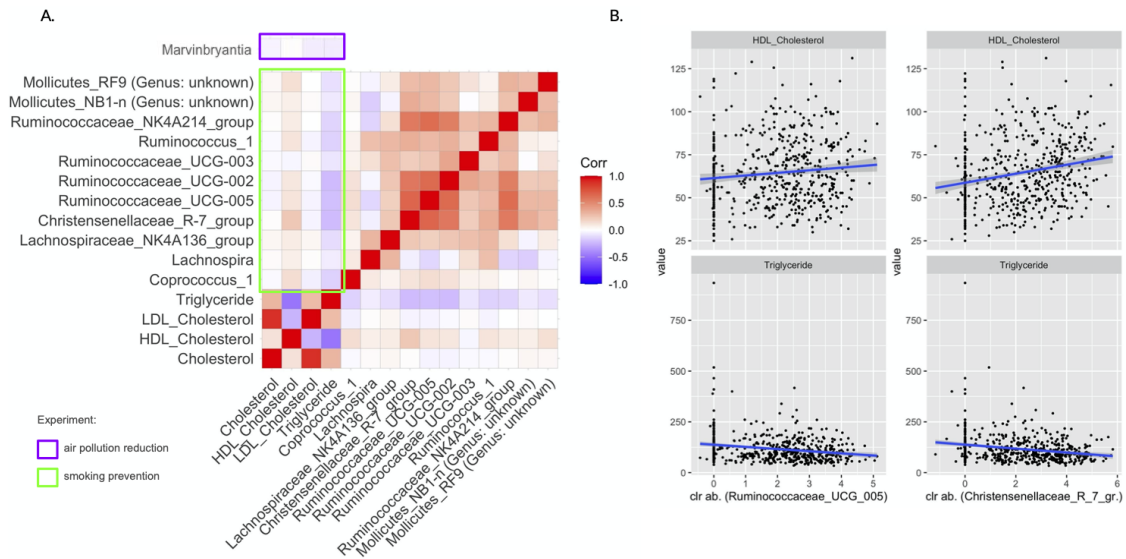


Figure 5: (A) Lipid metabolites correlation with selected genera from the smoking prevention experiment (green). (B) Scatterplots of high-density lipoprotein (HDL) cholesterol and triglycerides vs. centered log-ratio transformed relative abundances of the genera *Ruminococcaceae-UCG-005* and *Christensenellaceae-R-7-group*.

For instance, in the smoking prevention dataset, we observed a positive correlation of *Christensenellaceae R-7-group* and *Ruminococcaceae (UCG-005)* genus abundances (under centered log-ratio transformation) with HDL cholesterol and negative correlation with triglyceride levels, respectively (see Fig 5B). Similar correlation patterns were also found for the other genera for whom we rejected the no differential abundance hypothesis (see second and forth column in Fig 5A). Our findings were also in line with recently reported correlation results in Vojinovic et al. [65] using the Dutch LifeLines-DEEP cohort [13] and the Rotterdam Study [14].

Sensitivity analysis

To assess whether the pair-matching strategy chosen for the design stage influenced the conclusions of this study, we conducted a sensitivity analysis (see Sensitivity Analysis section in S1 Text). For that, we implemented the more commonly-used propensity score matching algorithm [97] and obtained matched samples of: (i) 158 participants living in low $PM_{2.5}$ areas and 158 participants living in higher $PM_{2.5}$ areas, and (ii) 290 smokers and 290 never smokers (see Table D and Figs T-Y in S1 Text for the balance diagnostics). For both hypothetical randomized experiments, using propensity score matching at the design stage results in analyzing more matched samples. The microbial diversity analyses lead to the same conclusion for both experiments despite different design stages (see Fig Z and Tables E-F in S1 Text). Overall, we also observe small approximate Fisherian p-values after performing the propensity score matching, in the same way we observe small approximate Fisherian p-values with our pair-matching strategy. The test statistics have the same direction and magnitude. For the air pollution reduction experiment, the adjusted p-values are higher when performing propensity score matching when checking for differential abundances, i.e., we cannot reject the sharp null hypothesis of no differential abundance for the *Marvinbryantia* genus. For the smoking prevention experiment, we can reject the sharp null of no differential abundance for the same taxa and additional ones when performing propensity score matching compared to pair-matching (see Table C and Table G in S1 Text).

4 Discussion

We first discuss the results presented above, then elaborate on the statistical framework we used for our analyses, and suggest statistical and epidemiological extensions of our work.

In the air pollution (PM_{2.5}) reduction hypothetical experiment, we reject the sharp null hypotheses of no richness, no α -diversity, no β -diversity, and no high-dimensional mean differences. We also reject the no differential abundance hypothesis for the *Marvinbryantia* genus, and the no differential association hypothesis between: the *Succinivibrio* and *Slackia* genera, as well as the *Ruminiclostridium* and *Cloacibacillus* genera. Experiments exposing mice to PM_{2.5} resulted in mixed findings concerning difference in microbial richness and diversity. This might be due to regional differences in the chemical composition of PM_{2.5} as well as differences in the duration of exposure [29]. Thus far, only one human study estimated associations between PM_{2.5} exposure and the gut microbiome, and investigated the pathway of diabetes induction associated with PM exposure [28]. One of their key findings was that PM_{2.5} exposure reduced α -diversity (measured by Chao1 and Shannon indices), which is consistent with our observations.

In the smoking prevention hypothetical experiment, we rejected the sharp null hypotheses of no richness, no α -diversity, no β -diversity, and no high-dimensional mean differences. We also rejected the no differential abundance hypothesis for eleven genera (five of the *Ruminococcaceae* family, three of the *Lachnospiraceae* family, one of the *Christensenellaceae* family, and two of the *Mollicutes* class), and the no differential association hypothesis between the *Ruminiclostridium-6* and *Ruminococcaceae-UCG-010* genera, and between the *Ruminiclostridium-6* and *Christensenellaceae R-7-group* genera. Interestingly, the associations of *Ruminococcaceae-UCG-010* and *Christensenellaceae R-7-group* with *Ruminiclostridium-6* were also found to be worth further scrutiny. Their positive associations in the genus-genus network of smokers was absent in the genus-genus network of the never-smokers. The one study comparing the gut microbiome of smokers ($n = 203$) and never-smokers ($n = 288$) with similar sample size has a men-only study population [44]. They did not find any differences in α -diversity (measured with the Shannon index), whereas we conclude that α -diversity analyses are worth further scrutiny. Lee et al.'s PERMANOVA analyses for β -diversity differences, measured with Jaccard and weighted UniFrac distances, suggested differences. We reject the sharp null hypothesis at the between-subject differences analysis level. In their analysis of bacterial taxa on the phylum level, smokers had an increased proportion of Bacteroidetes with decreased Firmicutes and Proteobacteria compared with never-smokers. When we compare these phyla, we do not observe the same differences (see Fig S in S1 Text). Also, our compositional difference analyses do not result in the same set of differentially abundant genera that were reported by Lee et al. [44]. These conflicting findings could be due to the fact that their study was done on Korean men only. Nonetheless, it shows that there is a lack of knowledge on the effects of smoking on the human gut microbiome and that additional scientific investigations are necessary to make causal conclusions.

Throughout the extensive statistical analyses presented in this paper, we have tested sharp null hypotheses of no effect of an intervention on a wide range of gut microbiome outcomes, ranging from high-level microbial diversity estimates to differential genus-genus associations. To do so, we have performed randomization-based inference based on 10,000 permutations. This mode of inference has been motivated by two reasons: (i) it is difficult to postulate a joint model for the potential outcomes, and thereby provide an estimate of (and uncertainty around) a causal estimand, and (ii) it has been shown that using the actual randomization procedure that led to the observed data helps to report valid Fisher-exact p-values as opposed to p-values relying on approximating null randomization distributions [46]. As an example, in our mean difference analyses, we found some differences between the null randomization distribution of the test statistic when approximated by permuting the intervention assignment vector and when drawn from the approximating asymptotic distribution (see Figs J-K in S1 Text). A natural extension of this study would be to use a Neymanian or Bayesian mode of inference to tackle the same research questions. There, simulations should support evidence whether the approach can indeed recover the then estimated causal effects. Simulating microbiome data requires effort so that the common properties, such as compositionality and zero-inflation, can be preserved, but re-sampling approaches [98] and generative models [99] have been developed to achieve this end.

An important component of our randomization-based procedure is that the permutations of the intervention assignment vector conserves the matched-pair design of the hypothetical randomized experiment. This strategy has been advocated by Rubin [100] in the context of randomized trials, and more recently by Bind and Rubin [46] in the context of hypothetical randomized experiments, because assumptions on the underlying distribution of the data are not required. Only few R packages were built to perform randomization-based inference while conserving the design of the intervention assignment. Therefore, for every analysis in our study, we imported a matrix of 10,000 unique randomized intervention assignments to calculate our p-values (see https://github.com/AliceSommer/Causal_Microbiome_Tutorial for a reproducible example on the American Gut Data [16, 101]). Nonetheless, the `DACOMP` and `NetCoMi` R packages provide flexible functions enabling the calculation of randomization-based p-values for our study design to test sharp null hypotheses of no difference in taxa abundance and associations, respectively. We advocate for more development of such user-friendly software functions permitting flexibility and accountability of the design stage of observational studies. P-value adjustments for multiple comparison also follow a fully randomization-based procedure, while preserving the design of the experiment. The method has proven to be more powerful while maintaining the family-wise error rate [91].

Notice that when presenting our results, we never accepted alternative hypotheses but only rejected sharp nulls when unadjusted and adjusted p-values were small, i.e., indicating the hypotheses warrants further scrutiny [82]. In the field of microbiome data analysis, the terms differential abundance and associations are frequently used. Researchers report “differentially abundant” and “differentially associated” sets of taxa after testing sharp null hypotheses of no effect of an intervention. This terminology implicitly implies acceptance of the alternative hypotheses. However, when testing sharp null hypotheses we assess the amount of evidence against them in the observed data, which does not prove the alternative hypothesis to be true.

During the design stage, the outcome variable was ignored and only pre-exposure covariates were considered. The chosen balanced data is a sub-sample of units that can be used to estimate the effects of an intervention. Omitting the outcome data until the analysis avoids “model cherry-picking”, because the effect of the intervention is estimated once, after a successful design stage. Nonetheless, at the design stage, we can only consider the observed pre-exposure variables but the assignment mechanism could depend on unobserved pre-exposure variables. In gut microbiome studies, diet is often an unobserved confounder. For example, in this study, dietary intake data was collected for only 1,469/2,033 (i.e., 72%) participants. We verified balance in dietary intake for our balanced data subset (see Figs H-I in S1 Text). Even though we made sure that the observed potential confounding covariates are fairly balanced, there could still be imbalances in other unobserved background covariates, which could have an effect on our results. In such cases, Rosenbaum [102] has recommended to consider sensitivity analyses of how the Fisher-exact p-value would change, had the intervention assignment been plausibly different, see also Bind and Rubin [46]. Subject-matter knowledge on the probability of the binary exposure (i.e., smoking or air pollution) given the observed and unobserved background covariates should guide the plausible range of “sensitivity” p-values and the reason why they could deviate from the p-value calculated based on the assumed hypothetical intervention assignment. This idea provides material for an extension of the framework presented in this study.

The framework suggested in this paper facilitates a more transparent interpretation of results than standard approaches directly modeling the observed outcome. First, interpretation is only valid within the range of the background covariates of the study population in the respective hypothetical experiment (see their detailed characteristics in Table 4 and Figs B-I in S1 Text). The data do not provide direct information for the “unmatched” units. In addition to our pair-matching strategy, we conducted a sensitivity analysis using a propensity score matching algorithm at the design stage, which led to more matched pairs, and thus a broader range of background covariates values (see Table D in S1 Text). Both matching algorithms do not lead to conflicting results in the smoking prevention experiments. In the air pollution reduction experiment, only the differential abundance analysis does not lead to the same overall conclusion. At this stage, the researcher can decide between a larger number of units or more similar groups of units to compare. When designing our hypothetical experiment, we chose a pair-matching strategy, because it creates similar pairs of

participants based on subject-matter knowledge. For example, the number of females and males in the intervention and control groups is identical after pair-matching, whereas with propensity score matching, these numbers slightly differ (see Table 4 and Table D in S1 Text). Note that the matching algorithm considerations should be *a priori* specified before any statistical analysis is performed. Ideally, the design stage should be conducted by a statistician who is not involved in the subsequent statistical analysis stage. Second, the assumed assignment mechanism and underlying assumptions have to be clearly stated to obtain meaningful p-values. Standard approaches usually make strong assumptions (e.g., linearity), whose discussions are often neglected. Modeling the observed data and solely adjusting for confounders by including them in a regression, without a design stage, can be unreliable, especially when the pre-exposure covariates distributions of the control and intervention units are not similar. For instance, Cochran and Rubin [47], Heckman et al. [103], and Rubin [104] have shown that regression models can estimate biased treatment effects when the true relationship between the covariates and the outcome is not modeled accurately. Dehejia and Wahba have also shown that standard nonexperimental estimators such as regression are sensitive to the specification used in the regression [105]. This is another reason why we opted for an inference method that does not rely on parametric assumptions.

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 examine the effect of interventions and provide results that can directly contribute to policy recommendations. Until now, relationships between inhaled environmental exposures and the human gut microbiome were not examined with causal inference methods, so a first step to make advances in the field is to test, whether air pollution and smoking have no effect on the units of our study. If so, a potential next step would be to work with a dataset adequate for balancing covariates along different doses of the exposure such as suggested in [106] and estimate a causal dose-response in order to protect populations at risk.

In the smoking prevention experiment, the subset of genera retained at the differential abundance analysis step was linked to the serum markers triglycerides and high-density lipoprotein in previous studies [95, 96, 65]. In our data, we observe correlations between these genera and metabolites in the same direction than previously found by Vojinovic [65] (see Fig 5). Serum triglycerides and high-density lipoprotein play a role in metabolic syndrome, and associations between smoking and metabolic syndrome have also been found previously [107]. Therefore, we suggest further investigation on the pathway of cigarette smoke impacting the gut, which in turn has effects on circulating metabolites (and metabolic syndrome). A logical next step would be to apply our framework to other cohorts with similar amplicon data preprocessing and available pre-exposure covariates such as the Dutch LifeLines-DEEP [13] and Rotterdam Studies [14], and observe whether our results replicate.

Acknowledgements

We thank all KORA participants and technical assistants without whose contributions this study could not have been realized. We also thank Stefanie Peschel and Viet Tran for testing the code for the tutorial with the American Gut Data as well as Barak Brill for his support in the DACOMP implementation. The computations in this paper were run on the FASRC Odyssey cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

References

1. Wikoff WR, Anfora AT, Liu J, Schultz PG, Lesley SA, Peters EC, et al. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(10):3698–3703.
2. Visconti A, Le Roy CI, Rosa F, Rossi N, Martin TC, Mohnhey RP, et al. Interplay between the human gut microbiome and host metabolism. *Nature Communications*. 2019;10(1).
3. Belkaid Y, Hand T. Role of the Microbiota in Immunity and inflammation Yasmine. *Cell*. 2015;157(1):121–141.
4. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*. 2014;505(7484):559–563.
5. David La, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al. Host lifestyle affects human microbiota on daily timescales. *Genome Biology*. 2014;15(7):R89.
6. Langdon A, Crook N, Dantas G. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Medicine*. 2016;8(1).
7. Thursby E, Juge N. Introduction to the human gut microbiota. *Biochemical Journal*. 2017;474(11):1823–1836.
8. Marchesi JR, Adams DH, Fava F, Hermes GDA, Hirschfield GM, Hold G, et al. The gut microbiota and host health: a new clinical frontier. *Gut*. 2016;65(2):330–339.
9. Young VB. The role of the microbiome in human health and disease: an introduction for clinicians. *BMJ*. 2017;356.
10. Pace NR, Stahl DA, Lane DJ, Olsen GJ. The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. In: C MK, editor. *Advances in Microbial Ecology*. vol. 9. Boston, MA: Springer; 1986. p. 1–55.
11. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. *Nature*. 2007;449(7164):804–810.
12. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. *Cell*. 2014;159(4):789–799.
13. Scholtens S, Smidt N, Swertz MA, Bakker SJ, Dotinga A, Vonk JM, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *International Journal of Epidemiology*. 2015;44(4):1172–1180.
14. Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol*. 2017;32(9):807–850.
15. He Y, Wu W, Zheng HM, Li P, McDonald D, Sheng HF, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature Medicine*. 2018;24(10):1532–1535.
16. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*. 2018;3(3).
17. Holle R, Happich M, Löwel H, Wichmann H; MONICA/KORA Study Group. KORA - A Research Platform for Population Based Health Research. *Gesundheitswesen (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany))*. 2005;67(S 01):19–25.

18. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. *Curr Opin Gastroenterol.* 2015;31(1):69–75.
19. R uckerl R, Schneider A, Breitner S, Cyrus J, Peters A. Health effects of particulate air pollution: A review of epidemiological evidence. *Inhalation Toxicology.* 2011;23(10):555–592.
20. Huang C, Shi G. Smoking and microbiome in oral, airway, gut and some systemic diseases. *Journal of translational medicine.* 2019;17(1):225–225.
21. Bind MC, Rubin DB. Bridging observational studies and randomized experiments by embedding the former in the latter. *Statistical Methods in Medical Research.* 2019;28(7):1958–1978.
22. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal.* 2017;11(12).
23. Kaplan GG, Dixon E, Panaccione R, Fong A, Chen L, Szyszkowicz M, et al. Effect of ambient air pollution on the incidence of appendicitis. *Canadian Medical Association Journal.* 2009;181(9):591–597.
24. Ananthakrishnan AN, McGinley EL, Binion DG, Saeian K. Ambient air pollution correlates with hospitalizations for inflammatory bowel disease: an ecologic analysis. *Inflamm Bowel Dis.* 2011;17(5):1138–45.
25. Kaplan GG, Szyszkowicz M, Fichna J, Rowe BH, Porada E, Vincent R, et al. Non-specific abdominal pain and air pollution: a novel association. *PLoS One.* 2012;7(10):1–8.
26. Peters A. Epidemiology: Air pollution and mortality from diabetes mellitus. *Nature Reviews Endocrinology.* 2012;8(12):706.
27. Alderete TL, Jones RB, Chen Z, Kim JS, Habre R, Lurmann F, et al. Exposure to traffic-related air pollution and the composition of the gut microbiota in overweight and obese adolescents. *Environmental Research.* 2018;161:472–478.
28. Liu T, Chen X, Xu Y, Wu W, Tang W, Chen Z, et al. Gut microbiota partially mediates the effects of fine particulate matter on type 2 diabetes: Evidence from a population-based epidemiological study. *Environment International.* 2019;130.
29. Bailey MJ, Naik NN, Wild LE, Patterson WB, Alderete TL. Exposure to air pollutants and the gut microbiota: a potential link between exposure, obesity, and type 2 diabetes. *Gut Microbes.* 2020;11(5):1188–1202.
30. Fouladi F, Bailey MJ, Patterson WB, Sioda M, Blakley IC, Fodor AA, et al. Air pollution exposure is associated with the gut microbiome as revealed by shotgun metagenomic sequencing. *Environment International.* 2020;138:105604.
31. M oller W, H au inger K, Winkler-Heil R, Stahlhofen W, Meyer T, Hofmann W, et al. Mucociliary and long-term particle clearance in the airways of healthy nonsmoker subjects. *Journal of Applied Physiology.* 2004;97(6):2200–2206.
32. Beamish LA, Osornio-Vargas AR, Wine E. Air pollution: An environmental factor contributing to intestinal disease. *Journal of Crohn’s and Colitis.* 2011;5(4):279–286.
33. Mutlu EA, Engen PA, Soberanes S, Urich D, Forsyth CB, Nigdelioglu R, et al. Particulate matter air pollution causes oxidant-mediated increase in gut permeability in mice. *Particle and Fibre Technology.* 2011;8:19.
34. Kish L, Hotte N, Kaplan GG, Vincent R, Tso R, G anzle M, et al. Environmental particulate matter induces murine intestinal inflammatory responses and alters the gut microbiome. *PLoS One.* 2013;8(4):1–15.

35. Li R, Navab K, Hough G, Daher N, Zhang M, Mittelstein D, et al. Effect of exposure to atmospheric ultrafine particles on production of free fatty acids and lipid metabolites in the mouse small intestine. *Environ Health Perspectives*. 2015;123(1):34–41.
36. Mutlu EA, Comba IY, Cho T, Engen PA, Yazıcı C, Soberanes S, et al. Inhalational exposure to particulate matter air pollution alters the composition of the gut microbiome. *Environmental Pollution*. 2018;240:817–830.
37. Wang W, Zhou J, Chen M, Huang X, Xie X, Li W, et al. Exposure to concentrated ambient PM_{2.5} alters the composition of gut microbiota in a murine model. *Particle and Fibre Toxicology*. 2018;15(1):1–13.
38. Salim SY, Kaplan GG, Madsen KL. Air pollution effects on the gut microbiota. *Gut Microbes*. 2014;5(2):215–219.
39. Gui X, Yang Z, Li MD. Effect of Cigarette Smoke on Gut Microbiota: State of Knowledge. *Frontiers in Physiology*. 2021;12.
40. Calkins BM. A meta-analysis of the role of smoking in inflammatory bowel disease. *Digestive Diseases and Sciences*. 1989;34(12):1841–1854.
41. Cosnes J, Beaugerie L, Carbonnel F, Gendre J. Smoking cessation and the course of Crohn's disease: An intervention study. *Gastroenterology*. 2001;120(5):1093–1099.
42. Benjamin JL, Hedin CRH, Koutsoumpas A, Ng SC, McCarthy NE, Prescott NJ, et al. Smokers with Active Crohn's Disease Have a Clinically Relevant Dysbiosis of the Gastrointestinal Microbiota. *Inflammatory Bowel Diseases*. 2011;18(6):1092–1100.
43. Biedermann L, Zeitz J, Mwinji J, Sutter-Minder E, Rehman A, Ott SJ, et al. Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans. *PloS one*. 2013;8(3):e59260–e59260.
44. Lee SH, Yun Y, Kim SJ, Lee EJ, Chang Y, Ryu S, et al. Association between Cigarette Smoking Status and Composition of Gut Microbiota: Population-Based Cross-Sectional Study. *Journal of clinical medicine*. 2018;7(9):282.
45. Fisher RA. *The Design of Experiments*. Edinburgh: Oliver and Boyd; 1935.
46. Bind MAC, Rubin DB. When possible, report a Fisher-exact P value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences*. 2020;117(32):19151–19158.
47. Cochran WG, Rubin DB. Controlling Bias in Observational Studies: A Review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*. 1973;35(4):417–446.
48. Rubin DB. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics*. 1973;29(1):185–203.
49. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974;66(5):688–701.
50. Rubin DB. Inference and Missing Data. *Biometrika*. 1976;63(3):581–592.
51. Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association*. 1986;81(396):945–960.
52. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY, USA: Cambridge University Press; 2015.
53. Rubin DB. For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*. 2008;2(3):808–840.

54. Willis A, Bunge J, Whitman T. Improved detection of changes in species richness in high diversity microbial communities. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2017;66(5):963–977.
55. Willis AD, Martin BD. Estimating diversity in networked ecological communities. *Biostatistics*. 2020;.
56. Cao Y, Lin W, Li H. Two-sample tests of high-dimensional means for compositional data. *Biometrika*. 2018;105(1):115–132.
57. Brill B, Amir A, Heller R. Testing for differential abundance in compositional counts data, with application to microbiome studies. *The Annals of Applied Statistics*. 2022;.
58. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*. 2015;11(5):e1004226.
59. Peschel S, Müller CL, von Mutius E, Boulesteix AL, Depner M. NetCoMi: network construction and comparison for microbiome data in R. *Briefings in Bioinformatics*. 2020;.
60. Sohn MB, Li H. Compositional mediation analysis for microbiome studies. *The Annals of Applied Statistics*. 2019;13(1):661–681.
61. Wang C, Hu J, Blaser MJ, Li H. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics*. 2019;36(2):347–355.
62. Sazal MR, Stebliankin V, Mathee K, Narasimhan G. Causal Inference in Microbiomes Using Intervention Calculus. *bioRxiv*. 2020;doi:10.1101/2020.02.28.970624.
63. Wade KH, Hall LJ. Improving causality in microbiome research: can human genetic epidemiology help? *Wellcome open research*. 2020;4:199–199.
64. Hughes D, Bacigalupe R, Wang J, Rühlemann M, Falony G, Joossens M, et al. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nature Microbiology*. 2020;5.
65. Vojinovic D, Radjabzadeh D, Kurilshikov A, Amin N, Wijmenga C, Franke L, et al. Relationship between gut microbiota and circulating metabolites in population-based cohorts. *Nature Communications*. 2019;10:5813.
66. Breuninger TA, Riedl A, Wawro N, Rathmann W, Strauch K, Quante A, et al. Differential associations between diet and prediabetes or diabetes in the KORA FF4 study. *Journal of Nutritional Science*. 2018;7:e34.
67. Godon JJ, Zumstein E, Dabert P, Habouzit F, Moletta R. Molecular microbial diversity of an anaerobic digester as determined by small-subunit rDNA sequence analysis. *Applied and environmental microbiology*. 1997;63(7).
68. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research*. 2013;41(1).
69. Callahan BJ, Sankaran K, Fukuyama JA, Mcmurdie PJ, Holmes SP. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses [version 1; referees: 2 approved]. *F1000Research*. 2016;5.
70. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*. 2013;41(Database issue):D590.

71. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*. 2007;73(16):5261.
72. Edgar RC, Valencia A. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*. 2018;34(14):2371–2375.
73. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*. 2005;71(12):8228–8235.
74. Wright ES. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal*. 2016;8(1):352–359.
75. Studier JA, Keppler KJ. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular biology and evolution*. 1988;5(6):729.
76. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*. 2007;26(1):20–36.
77. Micali S, Vazirani VV. An Algorithm for Finding Maximum Matching in General Graphs. In: *Proceedings of the 21st Annual Symposium on Foundations of Computer Science. SFCS '80*. Washington, DC, USA: IEEE Computer Society; 1980. p. 17–27.
78. Singh RK, Chang HW, Yan D, Lee KM, Ucmak D, Wong K, et al. Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine*. 2017;15(1):73.
79. Johnson AJ, Zheng JJ, Kang JW, Saboe A, Knights D, Zivkovic AM. A Guide to Diet-Microbiome Study Design. *Frontiers in Nutrition*. 2020;7:79.
80. Rubin DB. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*. 1980;75(371):591–593.
81. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*. 2017;8.
82. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond “ $p < 0.05$ ”. *The American Statistician*. 2019;73:1–19.
83. Willis A, Bunge J. Estimating diversity via frequency ratios. *Biometrics*. 2015;71(4):1042–1049.
84. Zhao N, Chen J, Carroll I, Ringel-Kulka T, Epstein M, Zhou H, et al. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *The American Journal of Human Genetics*. 2015;96(5):797–807.
85. Shannon CE. A Mathematical Theory of Communication. *Bell System Technical Journal*. 1948;27(3):379–423.
86. Basharin GP. On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. *Theory of Probability and its Applications*. 1959;4(3):333.
87. Brillinger DR, Jones LV, Tukey JW. The Role of Statistics in Weather Resources Management. In: *The Management of Weather Resources*. vol. 2. Washington D.C., USA: U.S. Government Printing Office; 1978. p. 25.
88. Aitchison JJ. *The statistical analysis of compositional data*. Caldwell, N.J.: Blackburn Press; 2003.
89. Gower JC. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*. 1971;27(4):857–871.

90. Lin X. Variance Component Testing in Generalised Linear Models with Random Effects. *Biometrika*. 1997;84(2):309–326.
91. Lee JJ, Forastiere L, Miratrix L, Pillai NS. More powerful multiple testing in randomized experiments with non-compliance. *Statistica Sinica*. 2017;27(3):1319–1345.
92. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008;9(3):432–441.
93. Liu H, Roeder K, Wasserman L. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *Adv Neural Inf Process Syst*. 2010;24(2):1432–1440.
94. Wasserstein R, Lazar N. The ASA’s Statement on p-Values: Context, Process, and Purpose. *American Statistician*. 2016;70(2):129–131.
95. Fu J, Bonder MJ, Cenit MC, Tigchelaar EF, Maatman A, Dekens JAM, et al. The Gut Microbiome Contributes to a Substantial Proportion of the Variation in Blood Lipids. *Circulation research*. 2015;117(9):817–824.
96. He Y, Wu W, Wu S, Zheng HM, Li P, Sheng HF, et al. Linking gut microbiota, metabolic syndrome and economic status based on a population-level analysis. *Microbiome*. 2018;6(1):172.
97. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
98. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*. 2013;8(4):1–11.
99. Ma S, Ren B, Mallick H, Moon YS, Schwager E, Maharjan S, et al. A statistical model for describing and simulating microbial community profiles. *PLOS Computational Biology*. 2021;17(9):1–27.
100. Rubin DB. More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine*. 1998;17(3):371–385.
101. Mishra AK, Müller CL. Negative Binomial factor regression with application to microbiome data analysis. *Statistics in Medicine*, accepted. 2022;.
102. Rosenbaum PR. *Design of Observational Studies*. Springer, New-York; 2010.
103. Heckman JJ, Ichimura H, Todd P. Matching as an econometric evaluation estimator. *Review of Economic Studies*. 1998;65:261–294.
104. Rubin DB. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology*. 2001;2(3):169–188.
105. Dehejia RH, Wahba S. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*. 1999;94(448):1053–1062.
106. 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. *Science advances*. 2020;6(29):eaba5692.
107. Sun K, Liu J, Ning G. Active Smoking and Risk of Metabolic Syndrome: A Meta-Analysis of Prospective Studies. *PLOS ONE*. 2012;7(10):e47791.

List of Supplementary Figures and Tables

Fig A in S1 Text: Gut microbiome data description. Number of observed ASV per sample (top left), sequencing depth per sample (top right), number of sequences per ASV (bottom left), number of zero count per ASV (bottom right).

Table A in S1 Text: Gut microbiome data description. Number of observed ASV per sample, sequencing depth per sample, number of sequences per ASV, number of zero count per ASV.

Fig B in S1 Text: Empirical distributions of the matched covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

Fig C in S1 Text: Empirical distributions of the disease covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

Fig D in S1 Text: Empirical distributions of the medication covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

Fig E in S1 Text: Empirical distributions of the matched covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

Fig F in S1 Text: Empirical distributions of the diseases covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

Fig G in S1 Text: Empirical distributions of the medication covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

Fig H in S1 Text: Empirical distributions of the nutrition covariates among the subjects under the intervention vs. not in the balanced data for the air pollution reduction hypothetical experiment.

Fig I in S1 Text: Empirical distributions of the nutrition covariates among the subjects under the intervention vs. not in the balanced data for the smoking prevention hypothetical experiment.

Fig J in S1 Text: Permutation-based (grey) and asymptotic (blue) null randomization distributions for the air pollution reduction hypothetical experiment.

Fig K in S1 Text: Permutation-based (grey) and asymptotic (blue) null randomization distributions for the smoking prevention hypothetical experiment.

Fig L in S1 Text: Reference set selection in the air pollution reduction experiment. A taxa enters the set $R = (r_1, \dots, r_F)$ if it has low variance (< 2) and high prevalence ($> 90\%$). For the analyses at the ASV level, we chose the variance to be < 3 and the prevalence to be $> 40\%$ as thresholds in order the have at least one reference per subject.

Fig M in S1 Text: Reference set selection in the smoking prevention experiment. A taxa enters the set $R = (r_1, \dots, r_F)$ if it has low variance (< 2) and high prevalence ($> 90\%$). For the analyses at the ASV level, we chose the variance to be < 3 and the prevalence to be $> 40\%$ as thresholds in order the have at least one reference per subject.

Fig N in S1 Text: Distribution of number of ASVs per sample when data is filtered at different ASV prevalence thresholds (0%, 5%, 10%, 15%) in the air pollution reduction experiment. Red value: minimum observed ASVs per sample.

Fig O in S1 Text: Distribution of the total ASV counts per sample when data is filtered at different ASV prevalence thresholds (0%, 5%, 10%, 15%) in the air pollution reduction experiment. Red value: minimum ASV counts per sample.

Fig P in S1 Text: Distribution of number of ASVs per sample when data is filtered at different ASV prevalence thresholds (0%, 5%, 10%, 15%) in the smoking prevention reduction experiment. Red value: minimum observed ASVs per sample.

Fig Q in S1 Text: Distribution of the total ASV counts per sample when data is filtered at different ASV prevalence thresholds (0%, 5%, 10%, 15%) in the smoking prevention experiment. Red value: minimum ASV counts per sample.

Table B in S1 Text: Air pollution reduction experiment results. Differentially abundant taxa and adjusted Fisher p-values for 10,000 iterations at 5% prevalence filtering. Selected adjusted p-values ≤ 0.2 (sign of abundance difference: $y(1) - y(0)$).

Table C in S1 Text: Smoking prevention experiment results. Differentially abundant taxa and adjusted Fisher p-values for 10,000 iterations at 5% prevalence filtering. Selected adjusted p-values ≤ 0.2 (sign of abundance difference: $y(1) - y(0)$).

Fig R in S1 Text: Genus-genus associations for subject under the air pollution reduction experiment vs. not ($n = 99$, $p = 149$). (A) Visualization of the between genera partial correlations estimated with the SPIEC-EASI method. Edges thickness is proportional to partial correlation, and color to direction: red: negative partial correlation, green: positive partial correlation. Node size is proportional to the centered log ratio of the genus abundances, and color is according to phyla. Triangle shaped nodes are differentially abundant (see Fig 3). (B) Zoom in largest connected component and differential associations (bold genera).

Fig S in S1 Text: Phyla comparison.

Table D in S1 Text: Sensitivity analysis - Baseline characteristics of the study population in the air pollution reduction (left table) and smoking prevention experiments (right table). Continuous variables: mean and standard deviation (St. d.). Categorical variables: number of samples per category (N) and proportion of category (%).

Fig T in S1 Text: Sensitivity analysis - Empirical distributions of the matched covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

Fig U in S1 Text: Sensitivity analysis - Empirical distributions of the diseases covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

Fig V in S1 Text: Sensitivity analysis - Empirical distributions of the medication covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

Fig W in S1 Text: Sensitivity analysis - Empirical distributions of the matched covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel)

data for the smoking prevention hypothetical experiment.

Fig X in S1 Text: Sensitivity analysis - Empirical distributions of the diseases covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

Fig Y in S1 Text: Sensitivity analysis - Empirical distributions of the medication covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

Fig Z in S1 Text: Sensitivity analysis - Richness and α -diversity. Boxplots (with median), values of the test-statistics from the **betta** regression, and one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

Table E in S1 Text: Sensitivity analysis - β -diversity. Microbiome Regression-based Kernel Association Test (**MiRKAT**), unadjusted and adjusted one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

Table F in S1 Text: Sensitivity analysis - Compositional equivalence test. Test statistic for high-dimensional data and one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

Table G in S1 Text: Sensitivity analysis - Smoking prevention experiment results. Differentially abundant taxa and adjusted Fisher p-values for 10,000 iterations at 5% prevalence filtering. Selected adjusted p-values ≤ 0.2 (sign of abundance difference: $y(1) - y(0)$).

Supplementary material: A randomization-based causal
inference framework for uncovering environmental exposure
effects on human gut microbiota

Alice J Sommer^{1,2,3,*}, Annette Peters^{2,3,4,*}, Martina Rommel^{3,5}, Josef Cyrus³, Harald
Grallert^{5,6}, Dirk Haller^{7,8}, Christian L Müller^{9,10,11,*}, and Marie-Abèle C Bind^{1,12}

¹Department of Statistics, Harvard University, Cambridge, MA, USA

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

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

⁴Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, MA, USA

⁵Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany

⁶German Center for Diabetes Research (DZD), München-Neuherberg, Germany

⁷ZIEL - Institute for Food & Health, Technical University of Munich, Freising, Germany

⁸Chair of Nutrition and Immunology, Technical University of Munich, Freising, Germany

⁹Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

¹⁰Department of Statistics, Ludwig-Maximilians-University München, Munich, Germany

¹¹Center for Computational Mathematics, Flatiron Institute, New York, NY, USA

¹²Biostatistics Center, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

*Corresponding authors: Alice J. Sommer: alice.j.sommer@gmail.com,

Annette Peters: peters@helmholtz-muenchen.de, and Christian L. Müller: cmueller@flatironinstitute.org

-

Gut microbiome data description (Amplicon Sequence Variants)

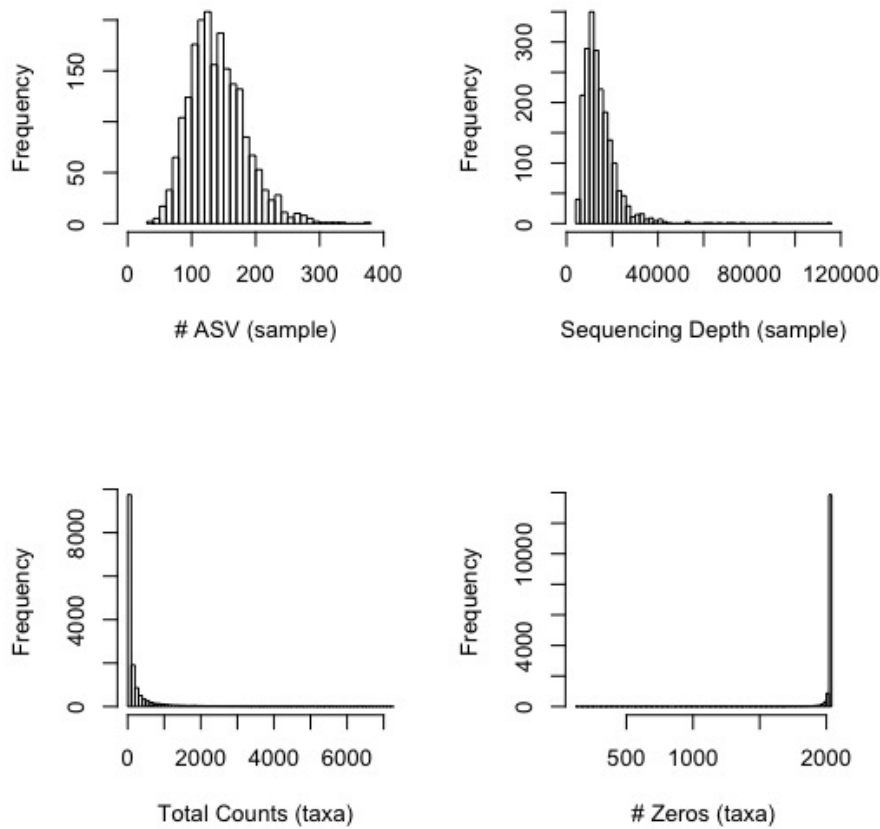


Fig A. Gut microbiome data description. Number of observed ASV per sample (top left), sequencing depth per sample (top right), number of sequences per ASV (bottom left), number of zero count per ASV (bottom right).

	Min.	1 st Qu.	Median	Mean	3 rd Qu.	Max.
nb. ASV (per sample)	31	109	135	140	168	371
nb. counts (per sample)	4,696	9,696	12,716	14,470	17,292	115,055
nb. counts (per taxa)	1	16	61	1,863	219	729,636
nb. zeros (per taxa)	122	2,030	2,033	2,016	2,033	2,033

Table A. Gut microbiome data description. Number of observed ASV per sample, sequencing depth per sample, number of sequences per ASV, number of zero count per ASV.

Balance diagnostics for matching

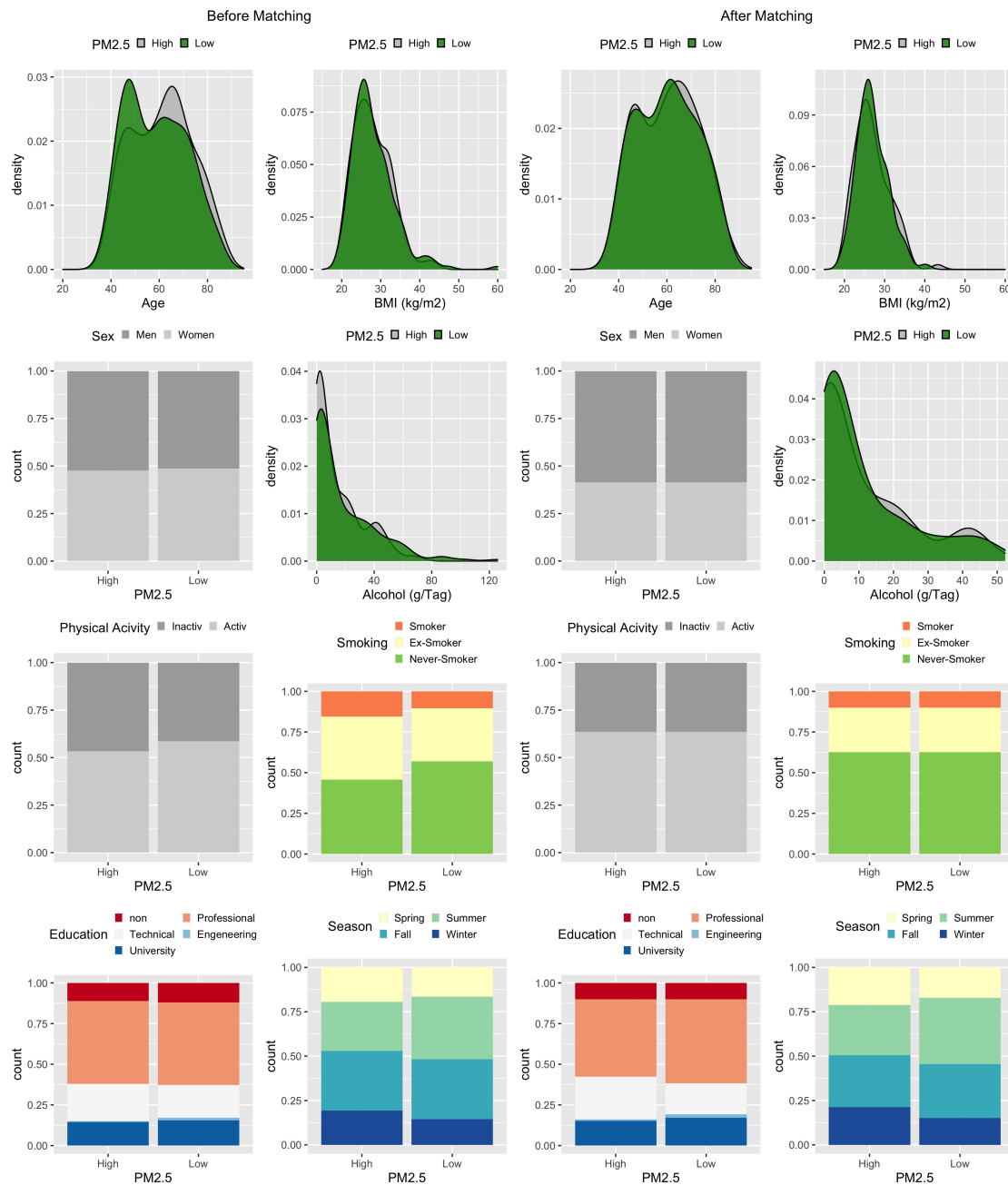


Fig B. Empirical distributions of the covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

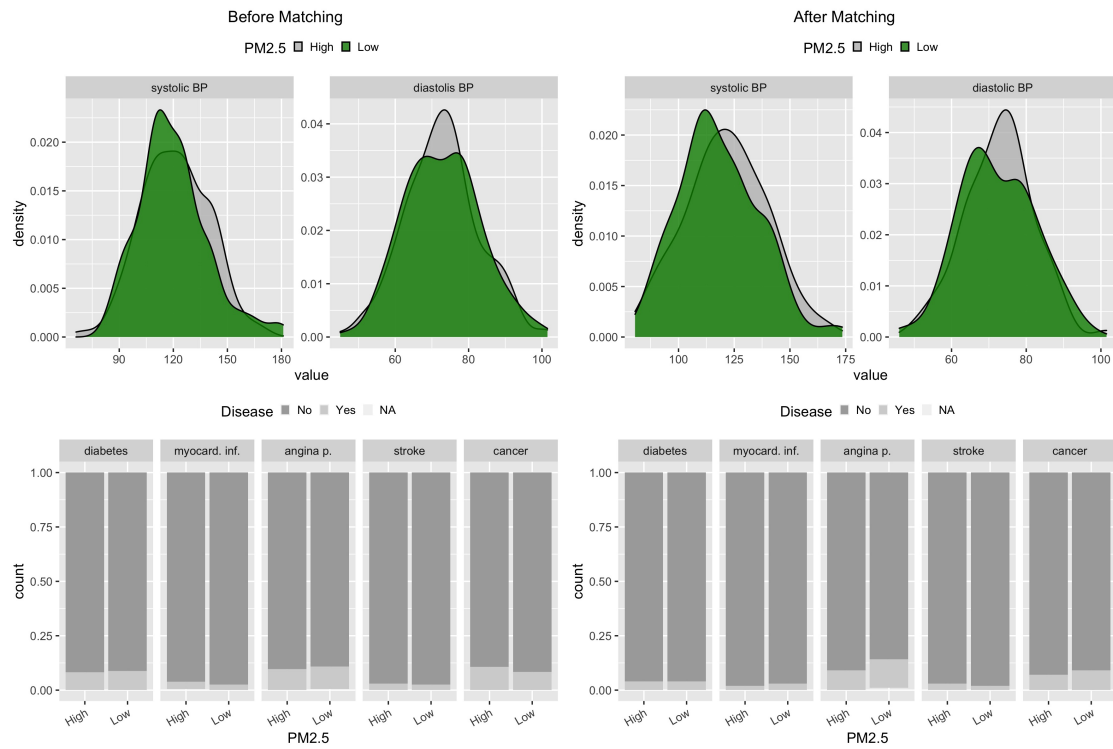


Fig C. Empirical distributions of the disease covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

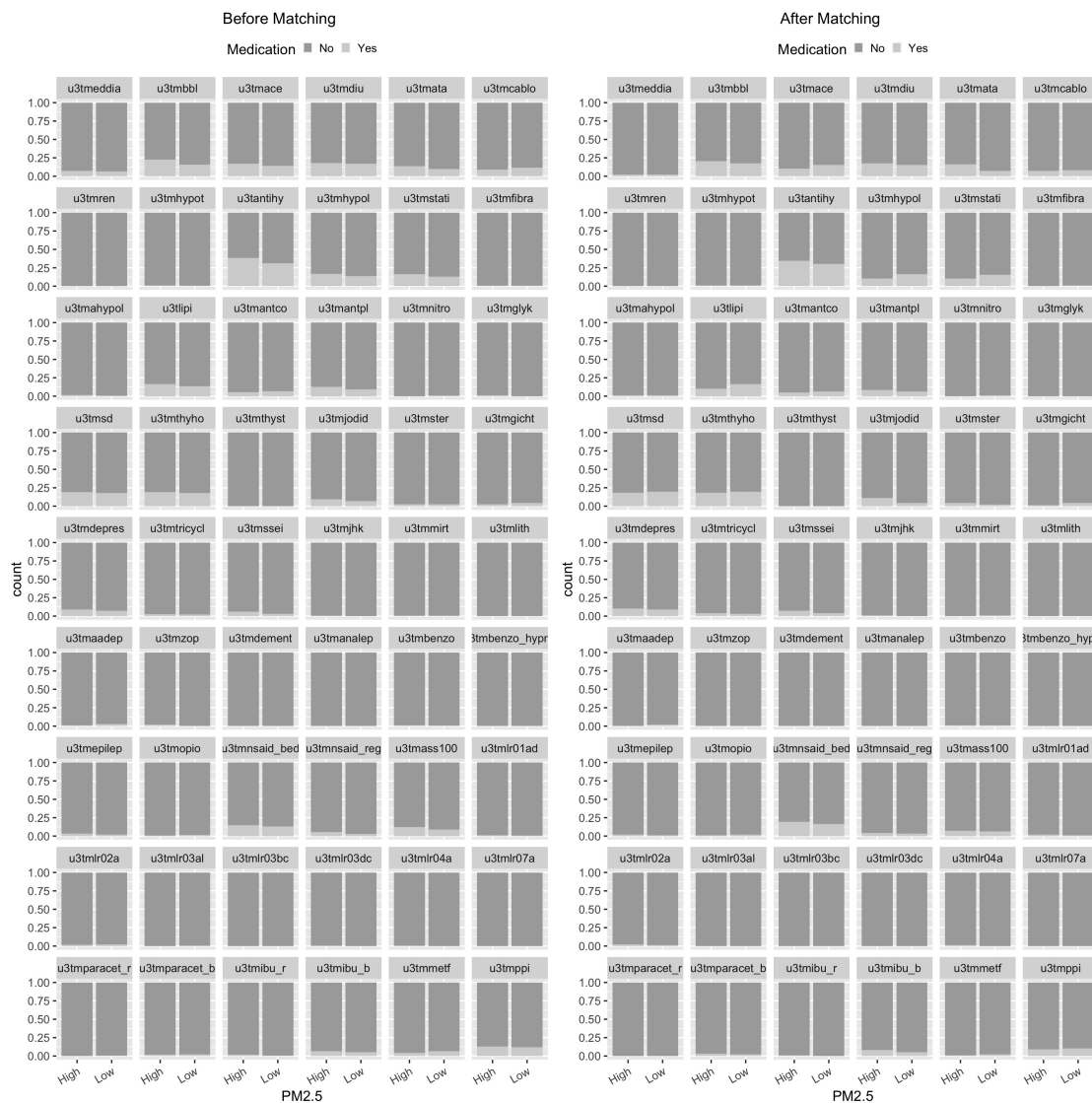


Fig D. Empirical distributions of the medication covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

Smoking

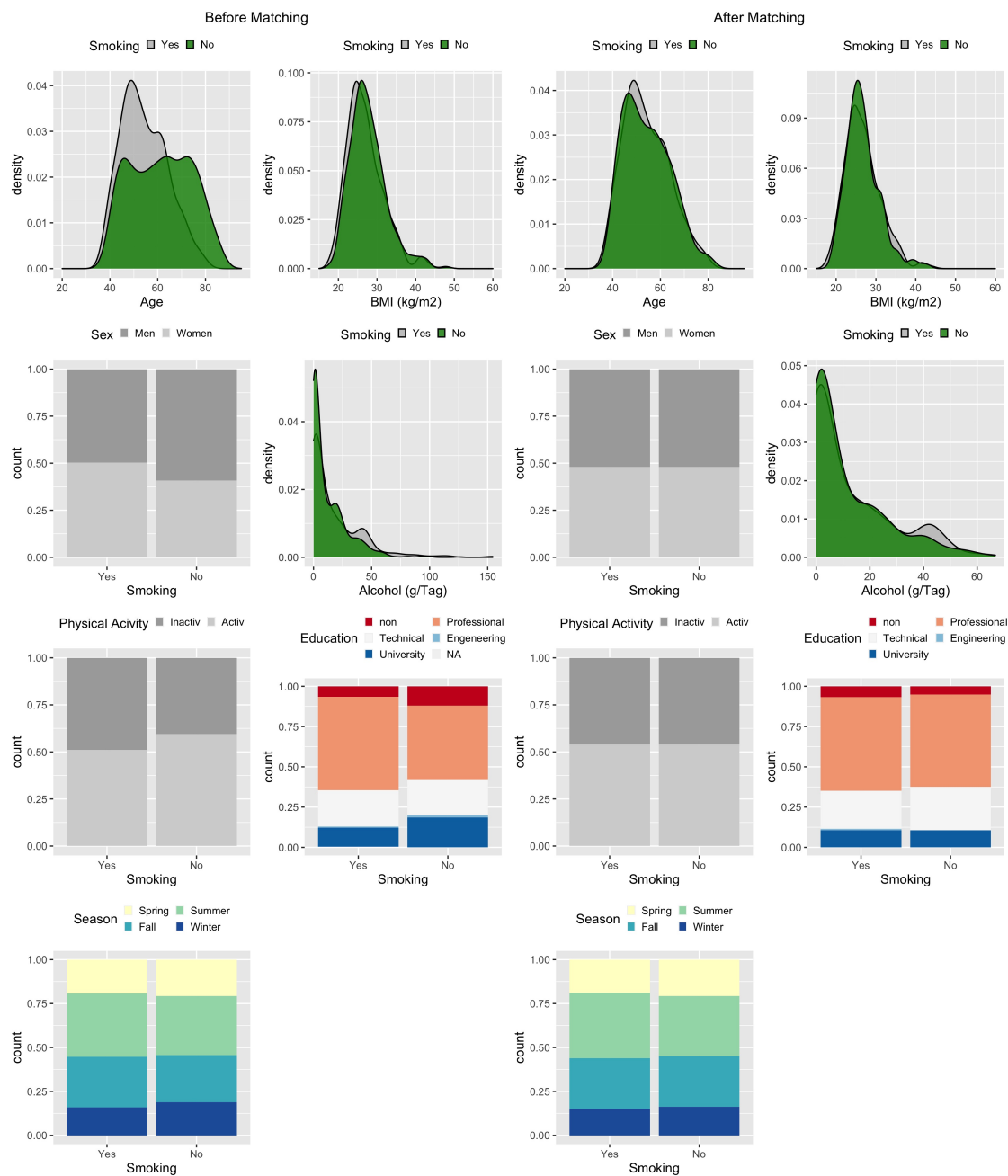


Fig E. Empirical distributions of the covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

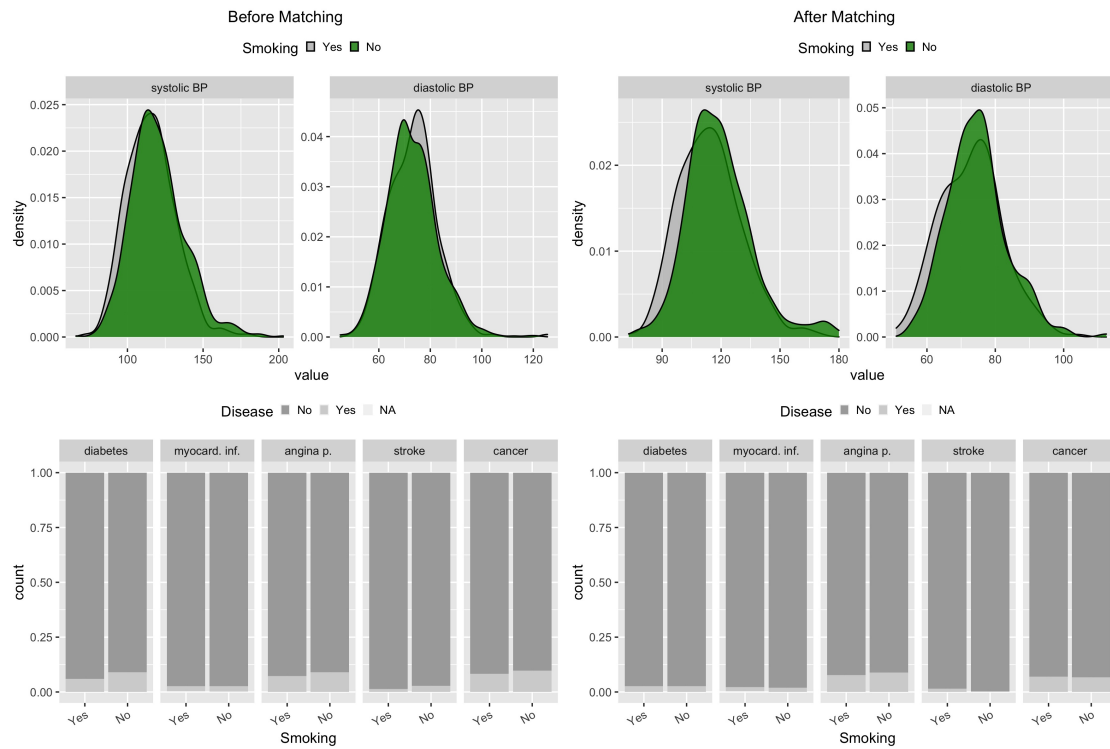


Fig F. Empirical distributions of the diseases covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

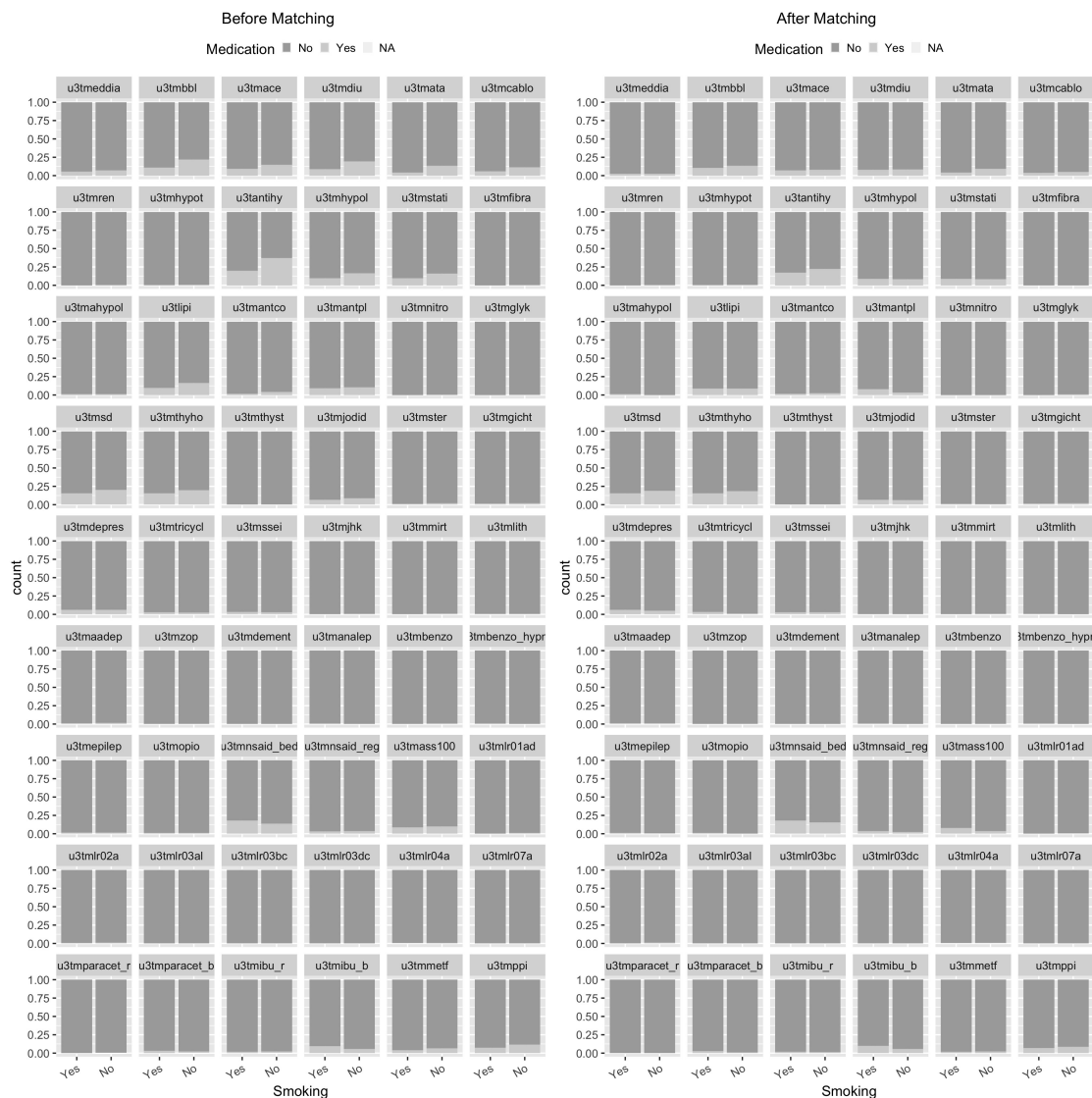


Fig G. Empirical distributions of the medication covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

Balance diagnostics for nutrition covariates after matching

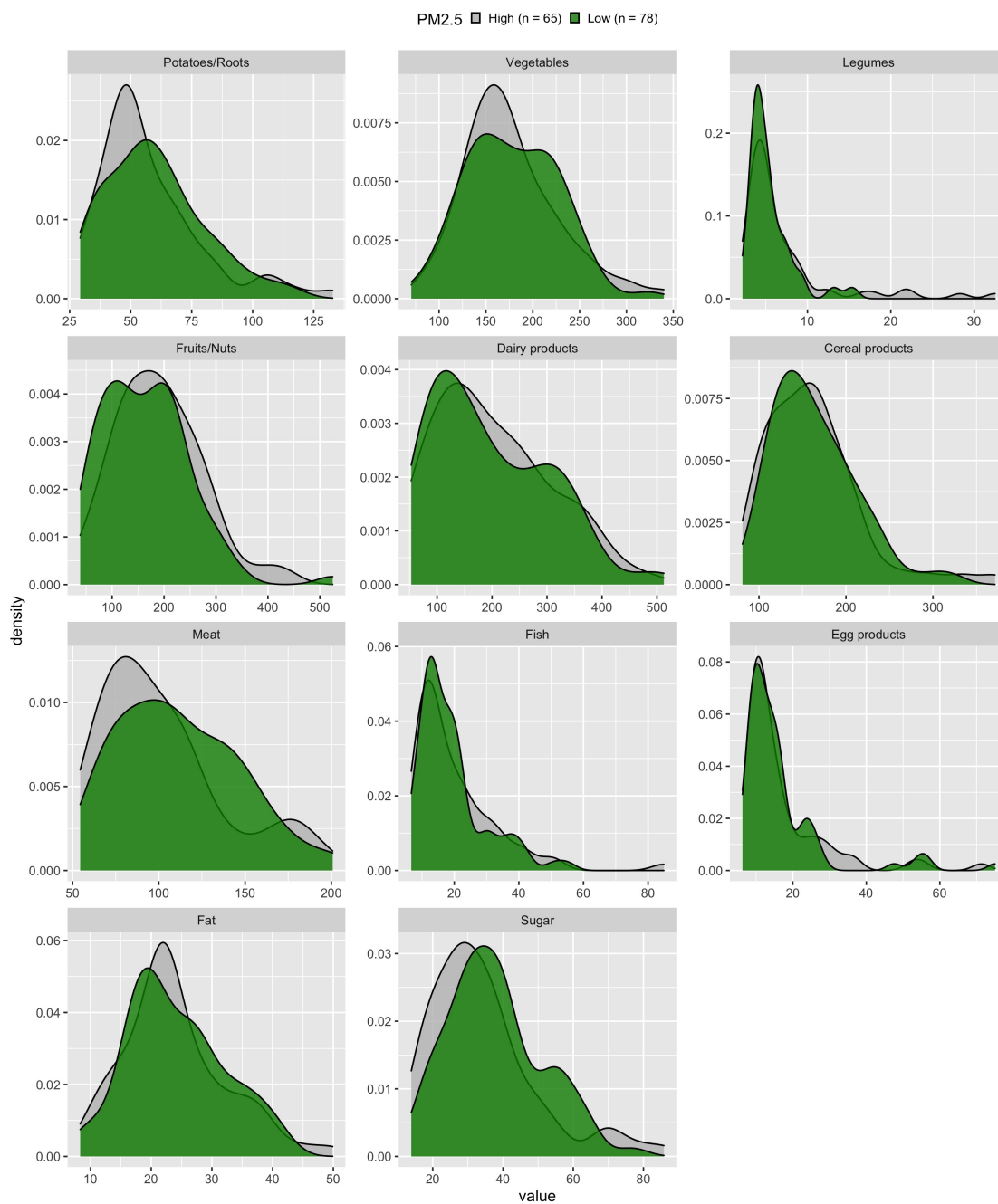


Fig H. Empirical distributions of the nutrition covariates among the subjects under the intervention vs. not in the balanced data for the air pollution reduction hypothetical experiment.

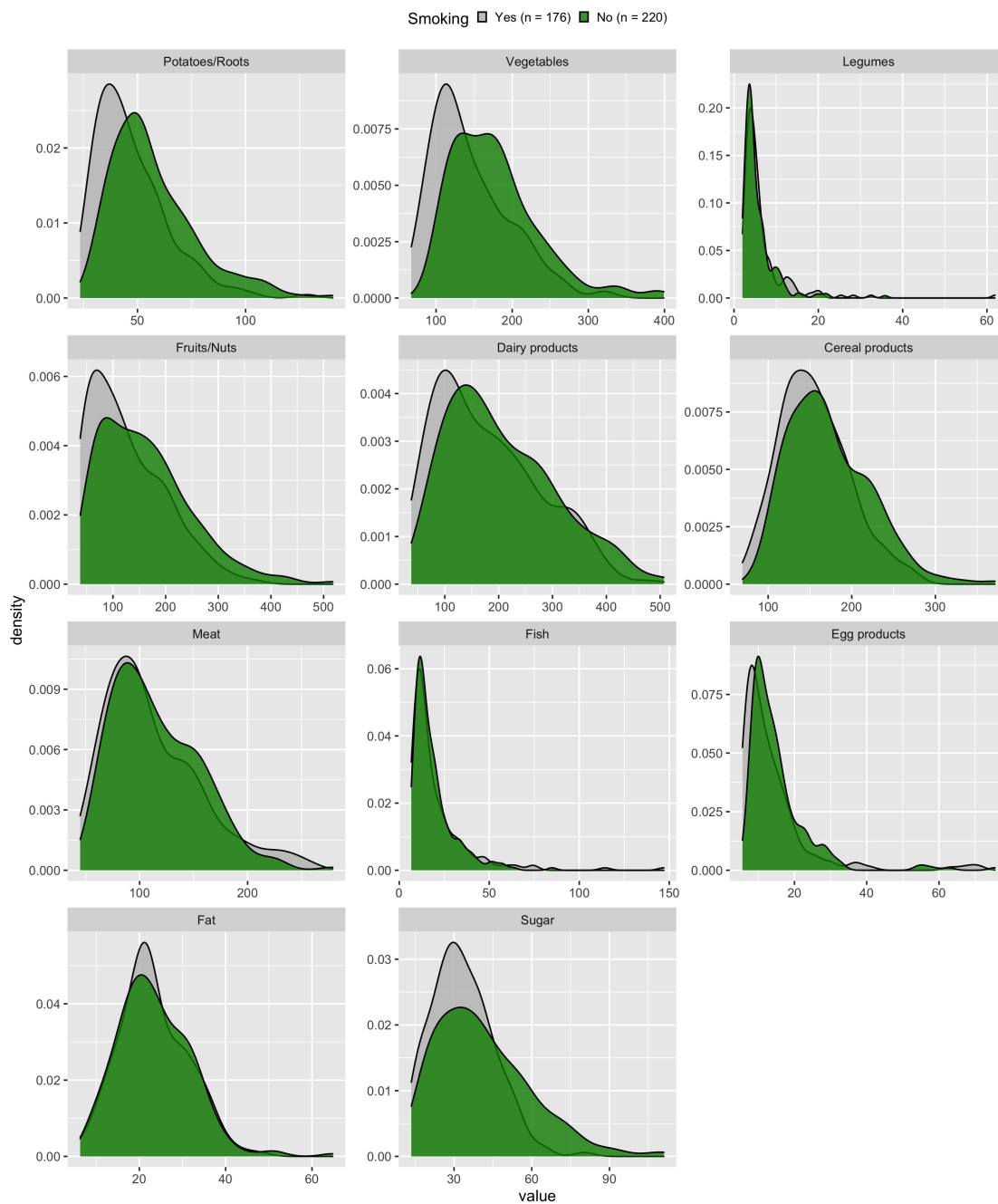


Fig I. Empirical distributions of the nutrition covariates among the subjects under the intervention vs. not in the balanced data for the smoking prevention hypothetical experiment.

Comparison of permutation and asymptotic null randomization distribution

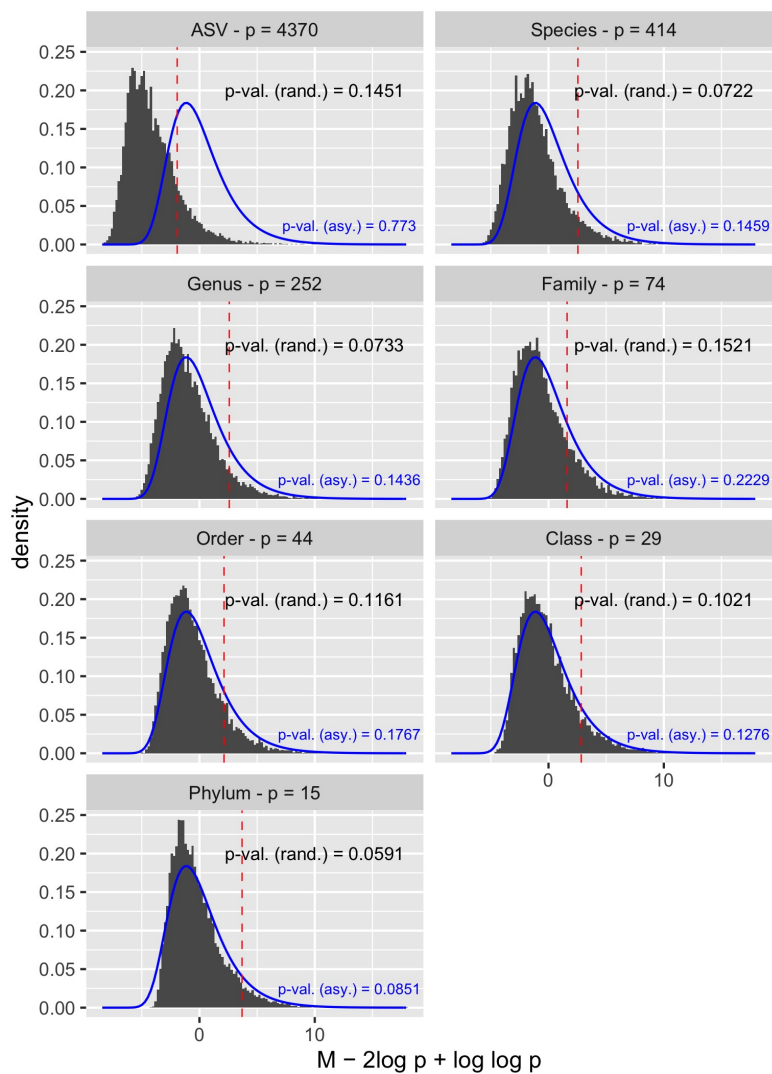


Fig J. Permutation-based (grey) and asymptotic (blue) null randomization distributions for the air pollution reduction hypothetical experiment.

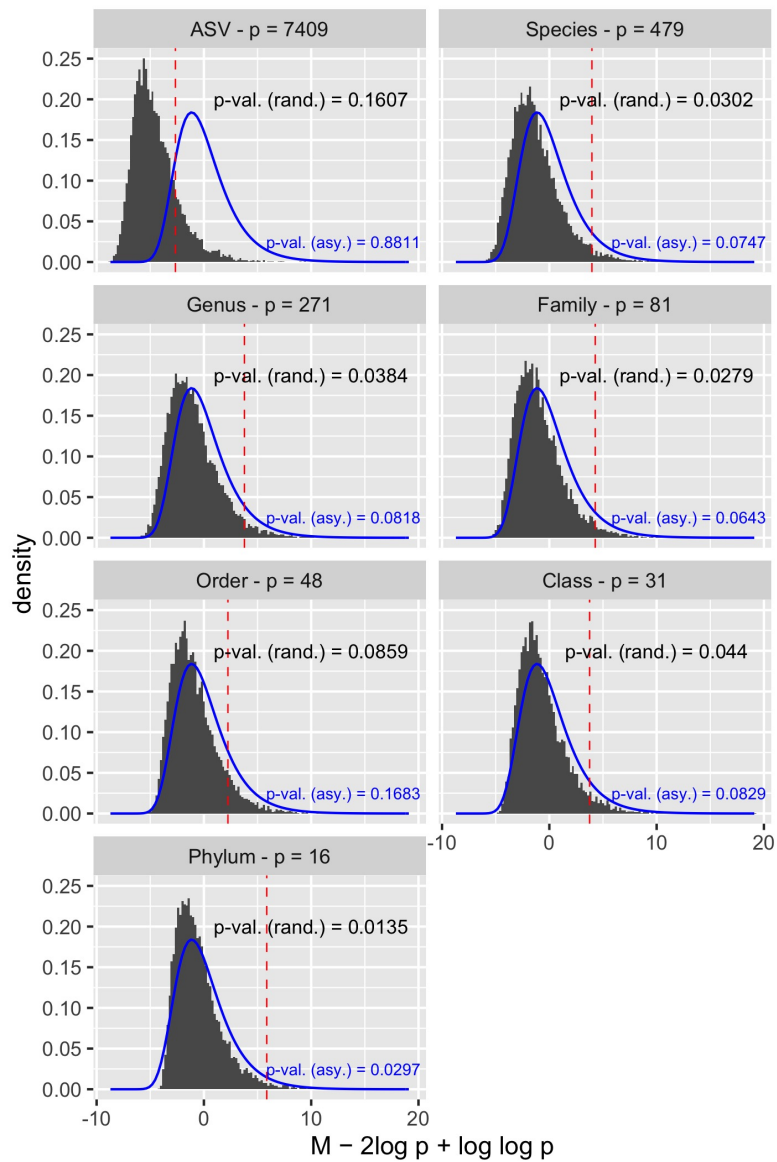


Fig K. Permutation-based (grey) and asymptotic (blue) null randomization distributions for the smoking prevention hypothetical experiment.

Reference selection for DACOMP

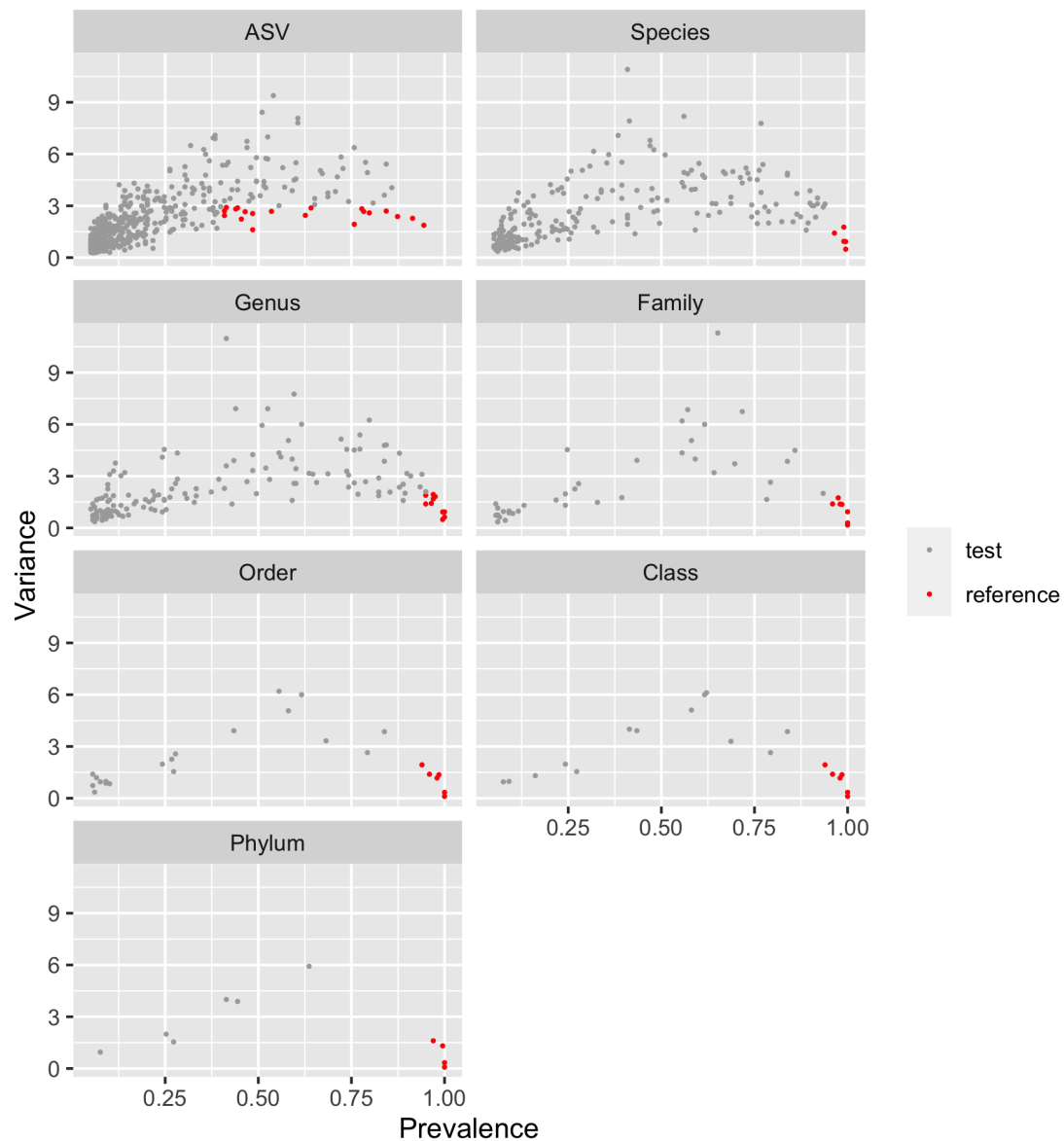


Fig L. Reference set selection in the air pollution reduction experiment. A taxa enters the set $R = (r_1, \dots, r_F)$ if it has low variance (< 2) and high prevalence ($> 90\%$). For the analyses at the ASV level, we chose the variance to be < 3 and the prevalence to be $> 40\%$ as thresholds in order to have at least one reference per subject.

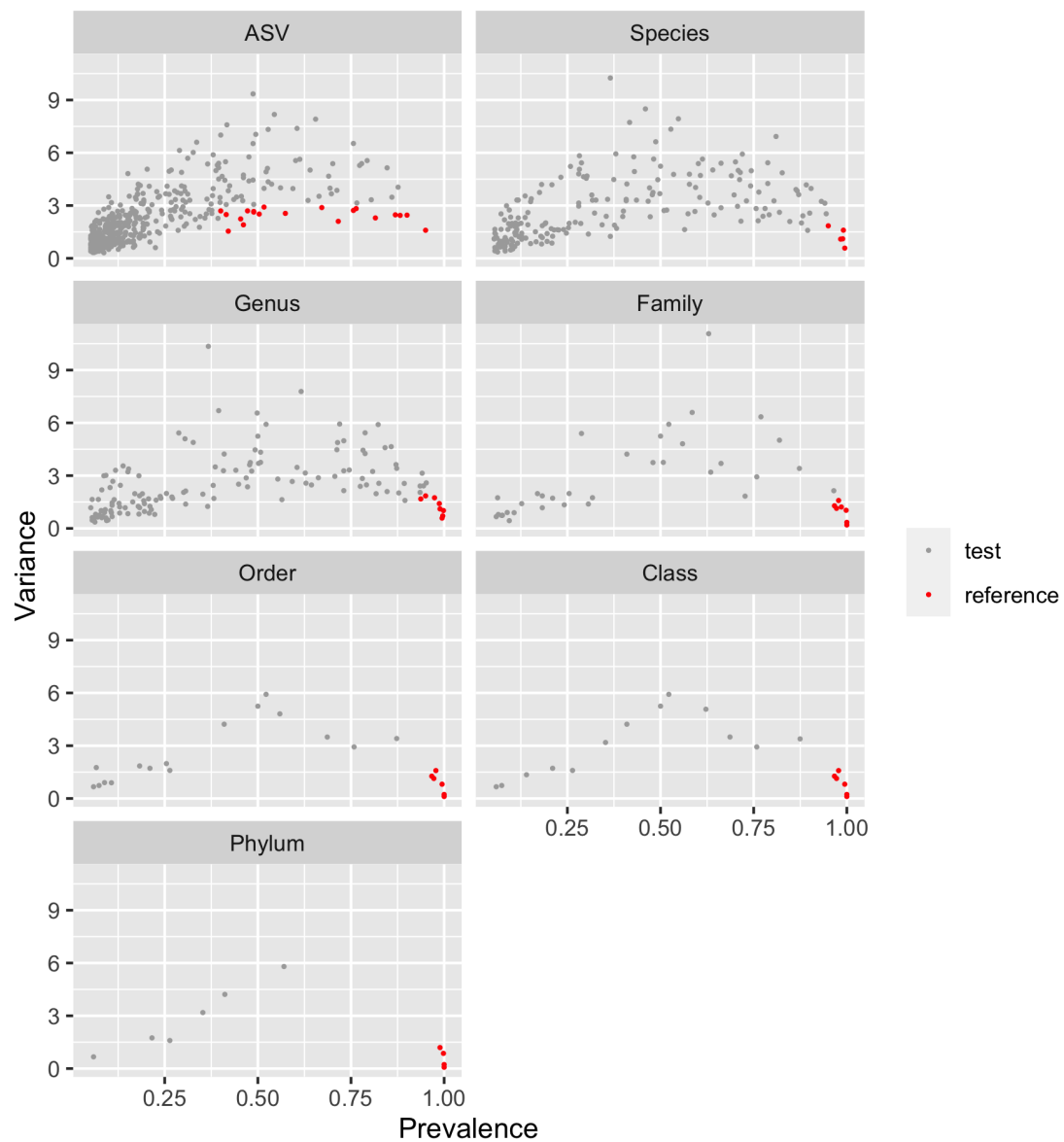


Fig M. Reference set selection in the smoking prevention experiment. A taxa enters the set $R = (r_1, \dots, r_F)$ if it has low variance (< 2) and high prevalence ($> 90\%$). For the analyses at the ASV level, we chose the variance to be < 3 and the prevalence to be $> 40\%$ as thresholds in order to have at least one reference per subject.

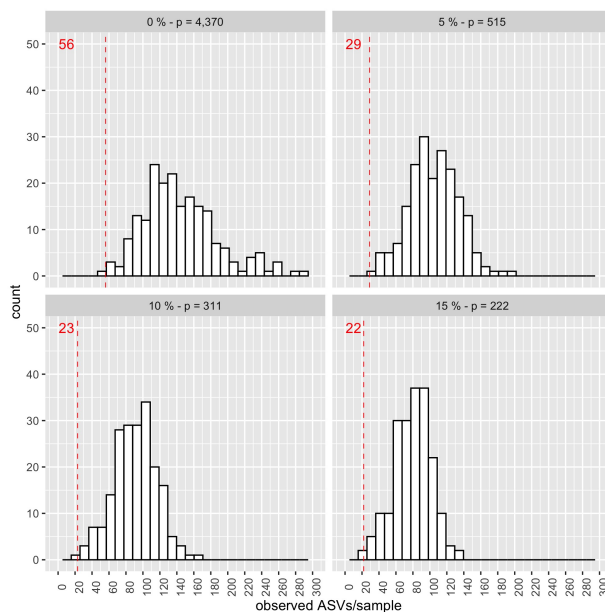


Fig N. Distribution of number of ASVs per sample when data is filtered at different ASV prevalence thresholds (0%, 5%, 10%, 15%) in the air pollution reduction experiment. Red value: minimum observed ASVs per sample.

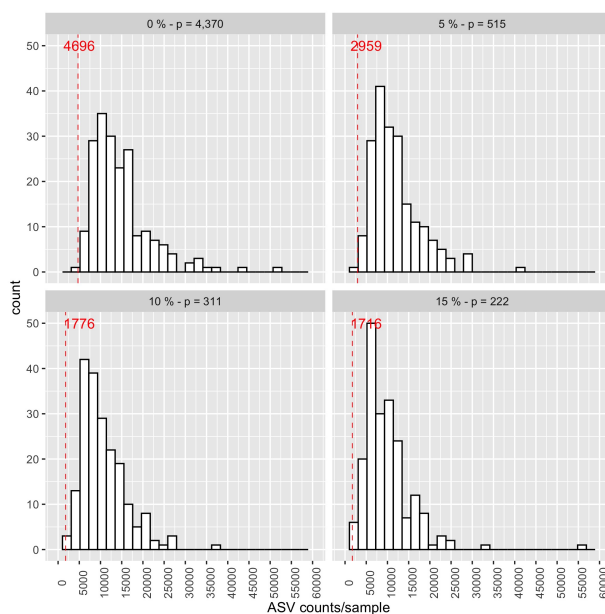


Fig O. Distribution of the total ASV counts per sample when data is filtered at different ASV prevalence thresholds (0%, 5%, 10%, 15%) in the air pollution reduction experiment. Red value: minimum ASV counts per sample.

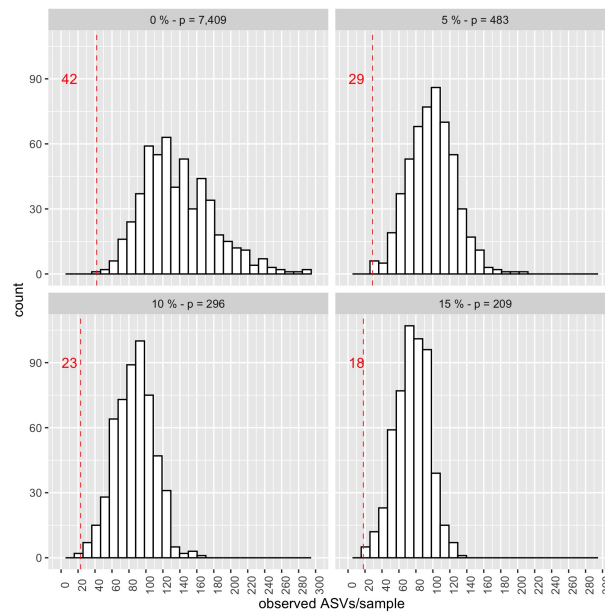


Fig P. Distribution of number of ASVs per sample when data is filtered at different ASV prevalence thresholds (0%, 5%, 10%, 15%) in the smoking prevention reduction experiment. Red value: minimum observed ASVs per sample.

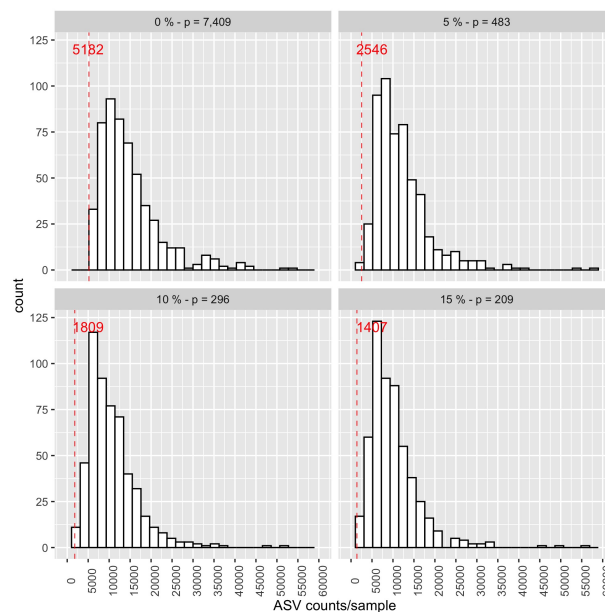


Fig Q. Distribution of the total ASV counts per sample when data is filtered at different ASV prevalence thresholds (0%, 5%, 10%, 15%) in the smoking prevention experiment. Red value: minimum ASV counts per sample.

	Kingdom	Phylum	Class	Order	Family	Genus	Species	p-value _{adj}
ASV p = 515	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus	NA	0.1461 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus	NA	0.0362 (-)
Species p = 220	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia	faecis	0.0357 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Marvinbryantia	NA	0.0181 (+)
Genus p = 149	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Marvinbryantia	NA	0.0120 (+)
Family p = 44								
Order p = 25								
Class p = 19								
Phylum p = 10								

Table B. Air pollution reduction experiment results. Differentially abundant taxa and adjusted Fisher p-values for 10,000 iterations at 5% prevalence filtering. Selected adjusted p-values ≤ 0.2 (sign of abundance difference: $y(1) - y(0)$).

	Kingdom	Phylum	Class	Order	Family	Genus	Species	p-value _{adj}
ASV p = 483	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus-1	NA	0.1250 (+)
Species p = 211	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-002	NA	0.1458 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae-NK4A136-group	NA	0.1124 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Christensenellaceae-R-7-group	NA	0.0201 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-005	NA	0.1124 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-003	NA	0.1297 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus-1	catus	0.0392 (+)
	Bacteria	Tenericutes	Mollicutes	NB1-n	NA	NA	NA	0.1458 (+)
	Bacteria	Tenericutes	Mollicutes	Mollicutes-RF9	NA	NA	NA	0.1791 (+)
Genus p = 140	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-002	NA	0.1476 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-003	NA	0.0127 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-005	NA	0.1975 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus-1	NA	0.1691 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-NK4A214-group	NA	0.1476 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Christensenellaceae-R-7-group	NA	0.0611 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospira	NA	0.0377 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae-NK4A136-group	NA	0.1781 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus-1	NA	0.0611 (+)
	Bacteria	Tenericutes	Mollicutes	NB1-n	NA	NA	NA	0.1882 (+)
	Bacteria	Tenericutes	Mollicutes	Mollicutes-RF9	NA	NA	NA	0.1166 (+)
	Family p = 41	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	NA	NA
Bacteria		Tenericutes	Mollicutes	NB1-n	NA	NA	NA	0.0450 (+)
Bacteria		Tenericutes	Mollicutes	Mollicutes-RF9	NA	NA	NA	0.0512 (+)
Order p = 22	Bacteria	Tenericutes	Mollicutes	NB1-n	NA	NA	NA	0.0375 (+)
	Bacteria	Tenericutes	Mollicutes	Mollicutes-RF9	NA	NA	NA	0.0404 (+)
Class p = 19	Bacteria	Tenericutes	Mollicutes	NA	NA	NA	NA	0.0039 (+)
Phylum p = 10	Bacteria	Tenericutes	NA	NA	NA	NA	NA	0.0018 (+)

Table C. Smoking prevention experiment results. Differentially abundant taxa and adjusted Fisher p-values for 10,000 iterations at 5% prevalence filtering. Selected adjusted p-values ≤ 0.2 (sign of abundance difference: $y(1) - y(0)$).



Fig R. Genus-genus associations for subject under the air pollution reduction experiment vs. not ($n = 99$, $p = 149$). (A) Visualization of the between genera partial correlations estimated with the SPIEC-EASI method. Edges thickness is proportional to partial correlation, and color to direction: red: negative partial correlation, green: positive partial correlation. Node size is proportional to the centered log ratio of the genus abundances, and color is according to phyla. Triangle shaped nodes are differentially abundant (see Figure 3). (B) Zoom in largest connected component and differential associations (bold genera).

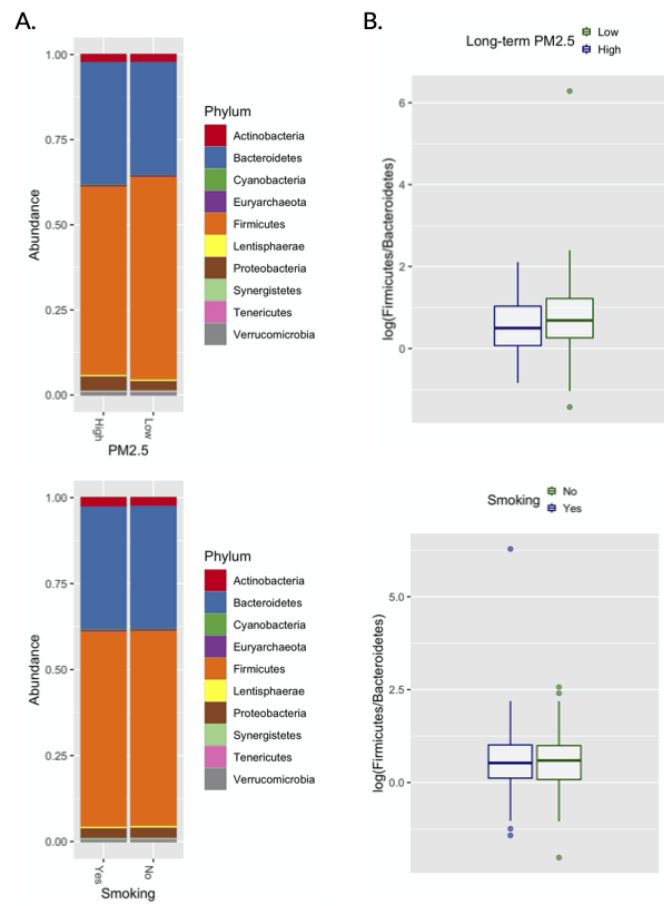


Fig S. Phyla comparison.

Sensitivity analysis

		Air pollution (PM _{2.5})				Smoking			
		≥ 13.0 μg/m ³ n = 158		≤ 10.3 μg/m ³ n = 158		Smoker n = 290		Never-Smoker n = 290	
		Mean	St. d.	Mean	St. d.	Mean	St. d.	Mean	St. d.
Age		60.0	12.8	59.6	12.6	54.6	9.4	54.8	11.4
Body Mass Index		28.1	5.6	28.1	5.5	27.0	4.9	27.0	4.5
Alcohol intake (g/day)		15.3	17.2	16.4	20.1	15.1	19.9	14.6	18.0
Years of education		11.8	2.8	11.4	2.5	11.7	2.4	12.0	2.6
		N	%	N	%	N	%	N	%
Sex	F	75	23.7	73	23.1	142	24.5	145	25.0
	M	83	26.3	85	26.9	148	25.5	145	25.0
Smoking	Ex-S.	60	19.0	57	18.0	-	-	-	-
	Never-S.	83	26.3	83	26.3	-	-	-	-
	Smoker	15	4.7	18	5.7	-	-	-	-
Diabetes	No	142	44.9	146	46.2	272	46.9	268	46.2
	Yes	16	5.1	12	3.8	18	3.1	22	3.8
Phys. Activity	No	70	22.2	66	20.9	137	23.6	127	21.9
	Yes	88	27.8	92	29.1	153	26.4	163	28.1

Table D. Sensitivity analysis - Baseline characteristics of the study population in the air pollution reduction (left table) and smoking prevention experiments (right table). Continuous variables: mean and standard deviation (St. d.). Categorical variables: number of samples per category (N) and proportion of category (%).

Air pollution

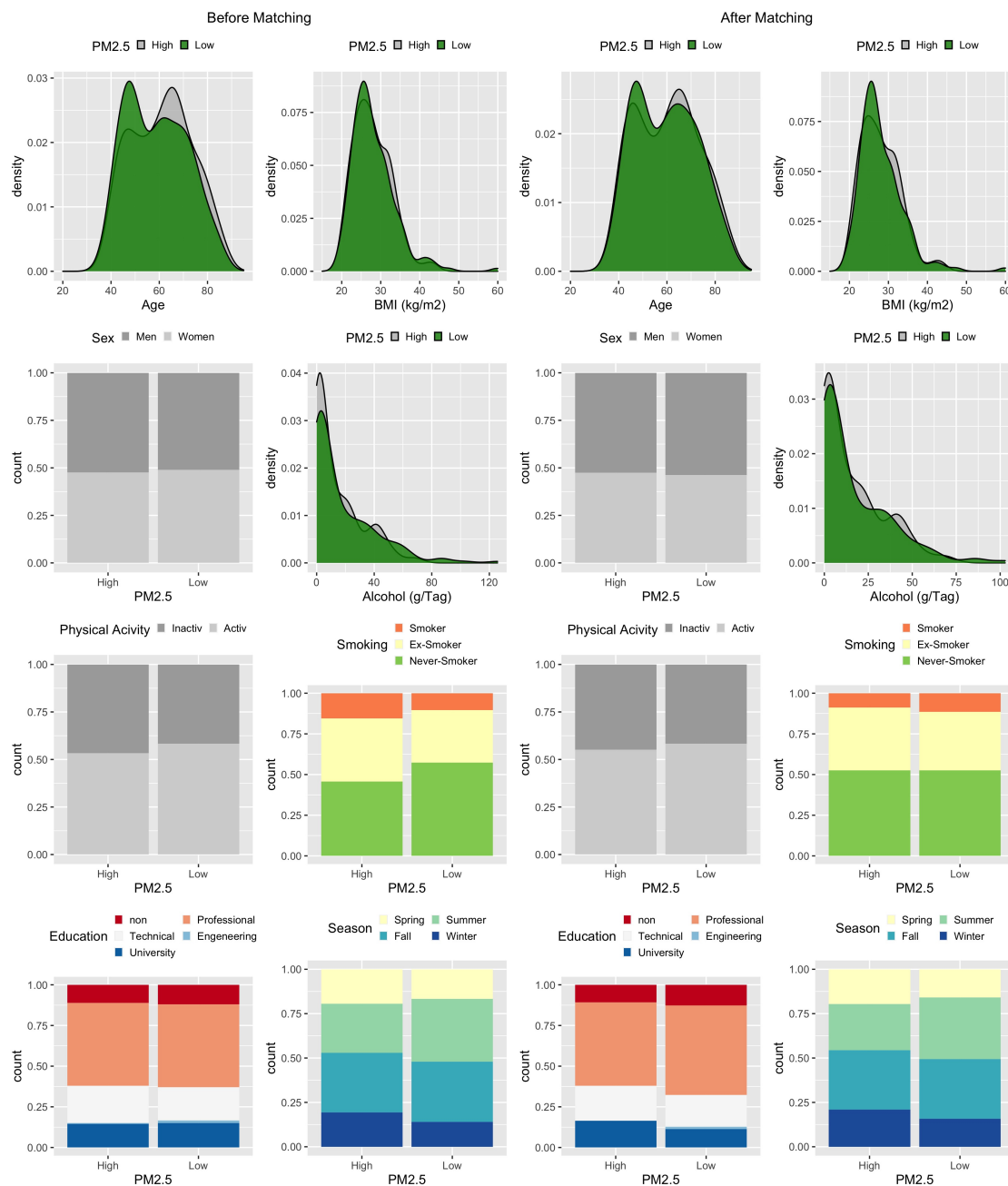


Fig T. Sensitivity analysis - Empirical distributions of the covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

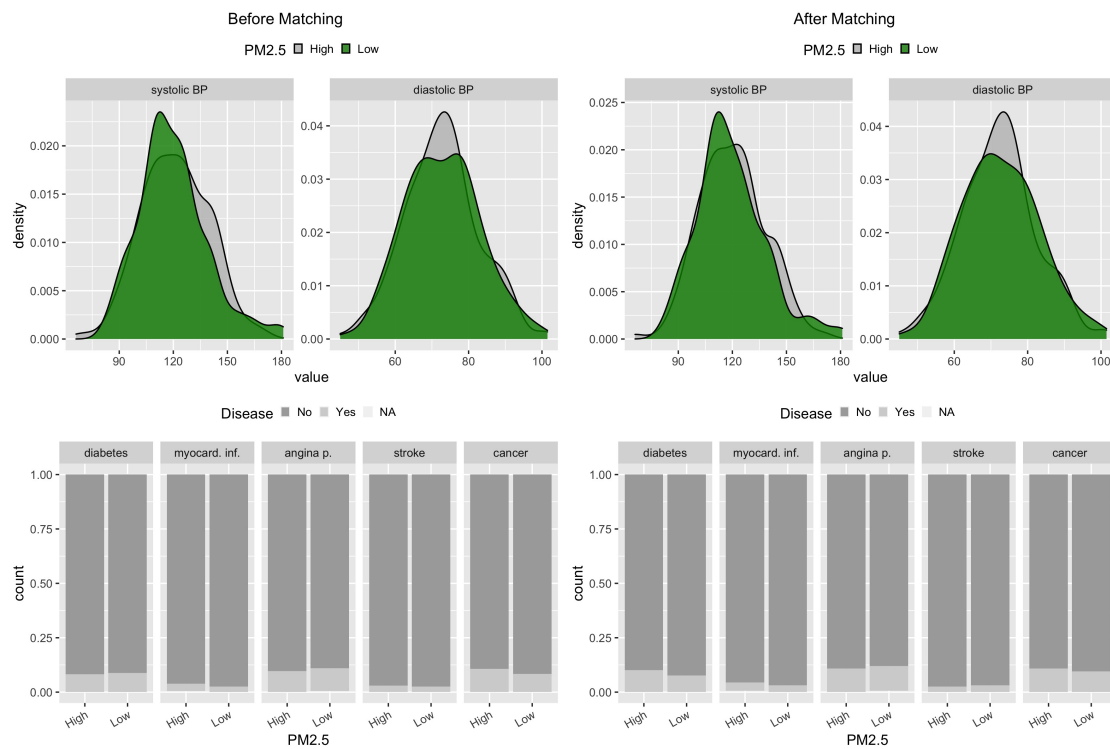


Fig U. Sensitivity analysis - Empirical distributions of the diseases covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

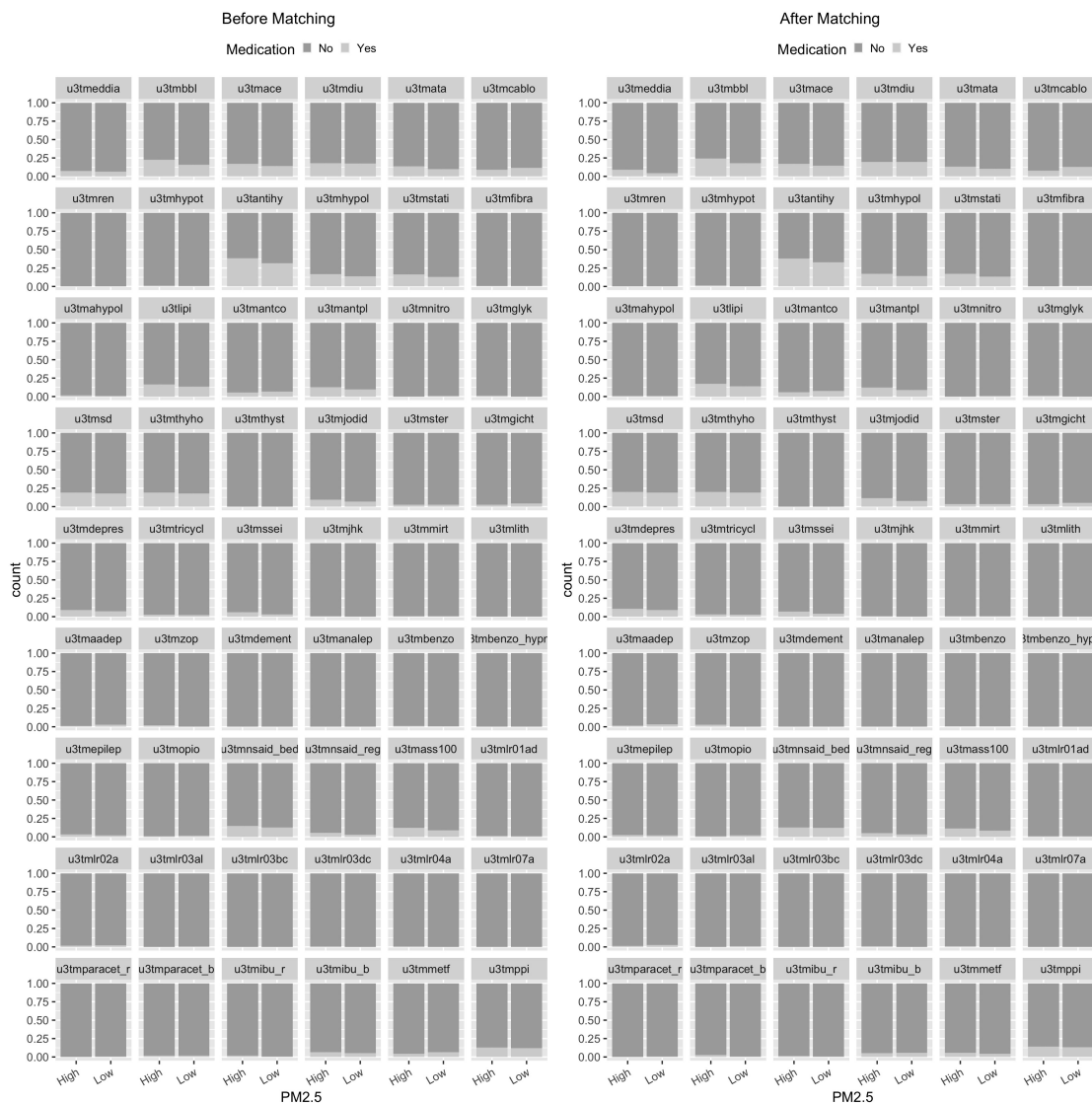


Fig V. Sensitivity analysis - Empirical distributions of the medication covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the air pollution reduction hypothetical experiment.

Smoking

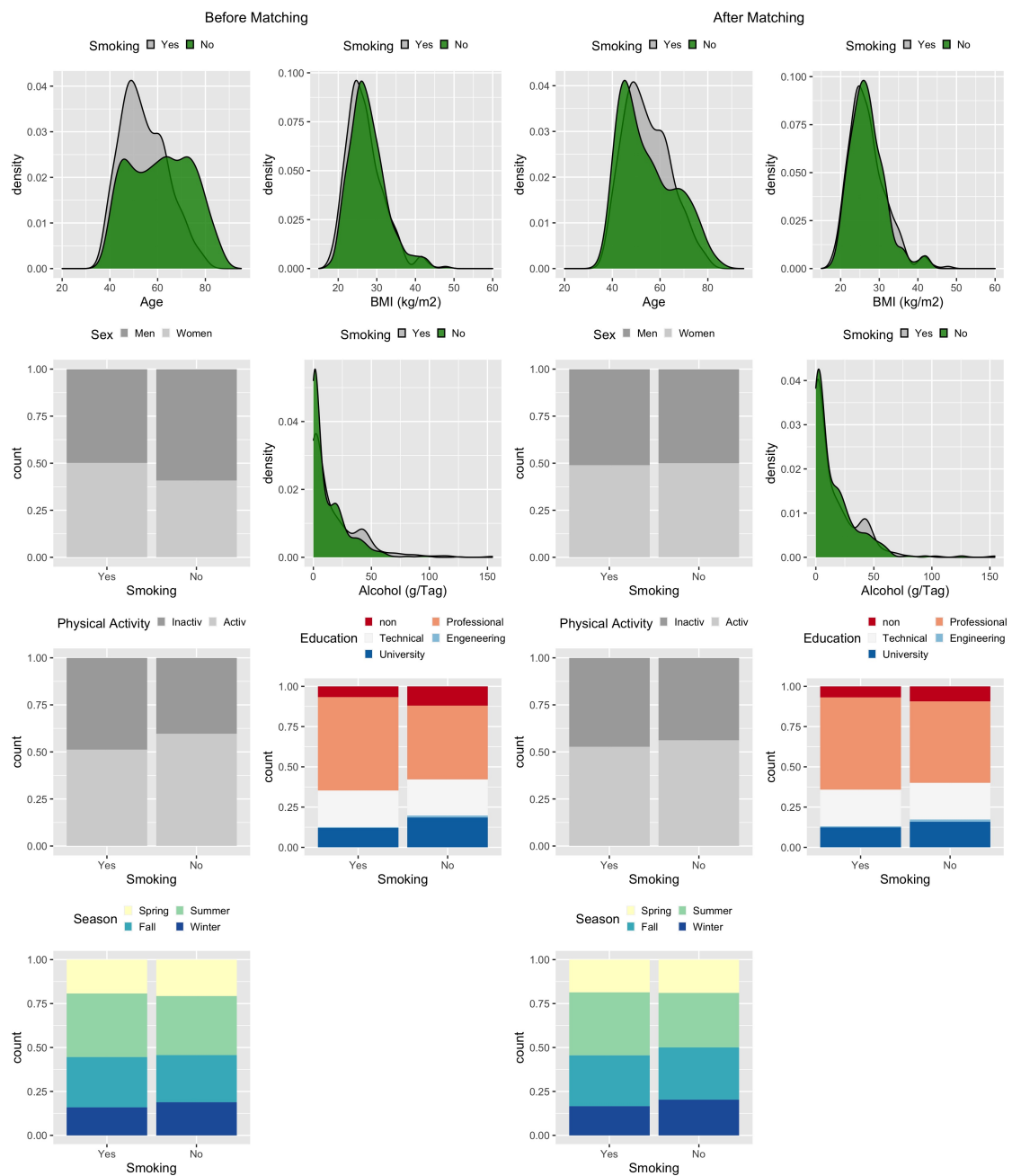


Fig W. Sensitivity analysis - Empirical distributions of the covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

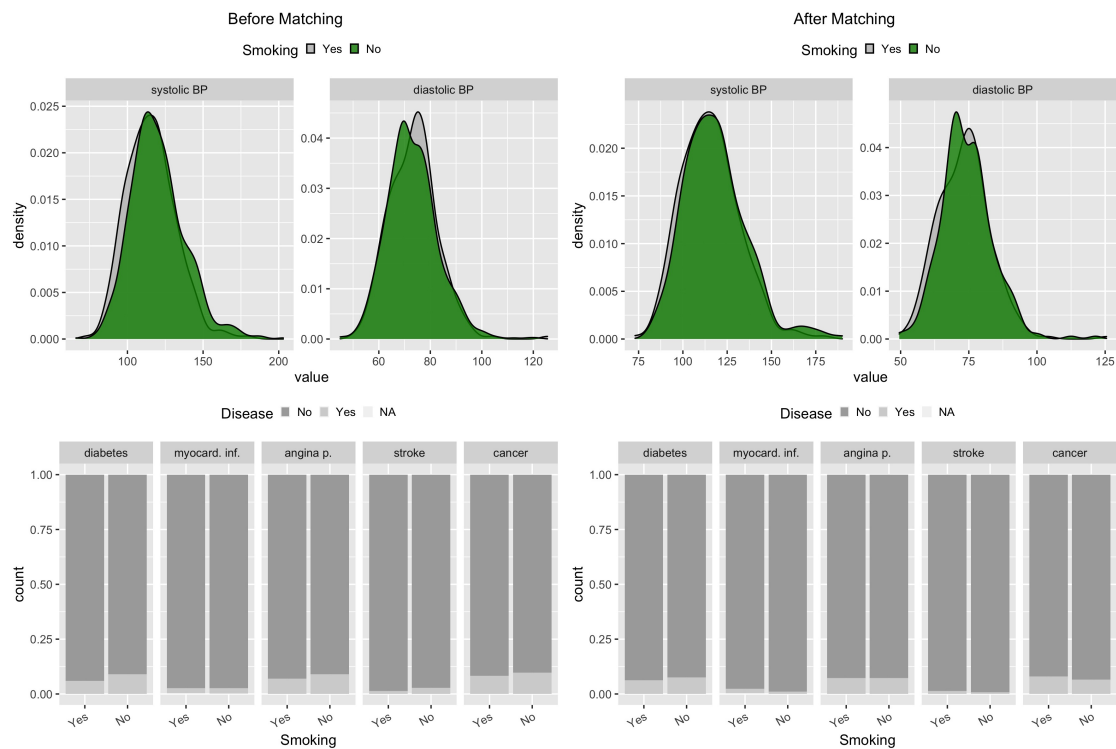


Fig X. Sensitivity analysis - Empirical distributions of the diseases covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

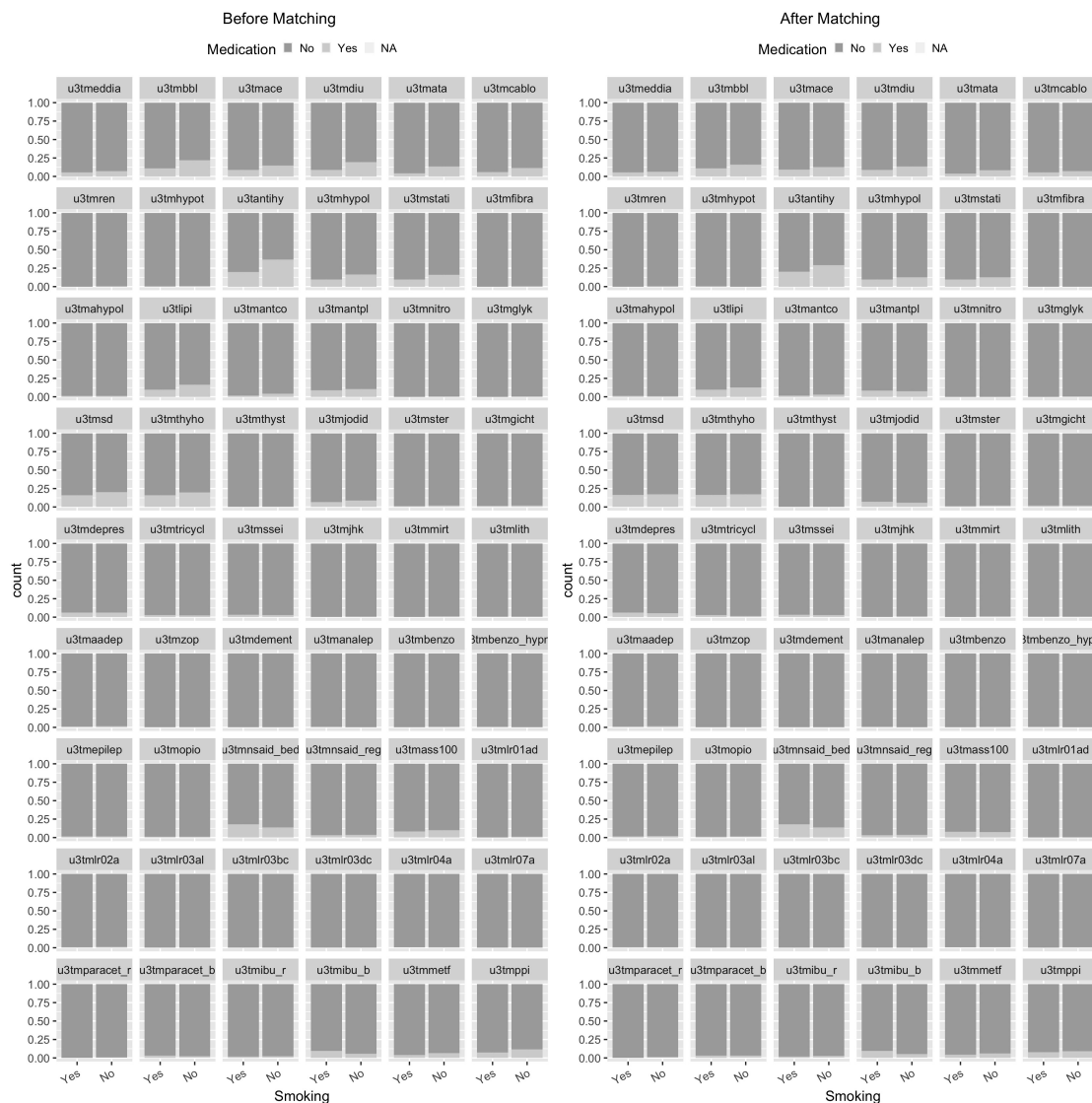


Fig Y. Sensitivity analysis - Empirical distributions of the medication covariates among the subjects under the intervention vs. not in the original (left panel) and the balanced (right panel) data for the smoking prevention hypothetical experiment.

Results

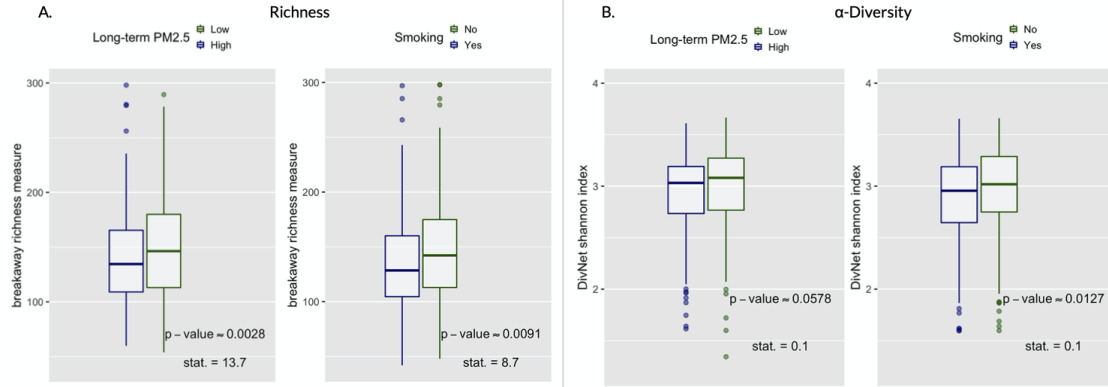


Fig Z. Sensitivity analysis - Richness and α -diversity. Boxplots (with median), values of the test-statistics from the **beta** regression, and one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

<i>distance</i>	Air pollution			Smoking		
	test-statistic	p-value	p-value _{adj}	test-statistic	p-value	p-value _{adj}
UniFrac	15.1	0.1950	0.3984	91.0	0.0004	0.0010
Aitchison	123180.7	0.2361	0.4658	432662.8	0.0003	0.0003
Jaccard	29.1	0.2238	0.4467	104.4	0.0001	0.0003
Gower	0.1	0.0223	0.0596	0.2	0.0046	0.0132

Table E. Sensitivity analysis - β -diversity. Microbiome Regression-based Kernel Association Test (MiRKAT), unadjusted and adjusted one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

		ASV	Species	Genus	Family	Order	Class	Phylum
Air Pollution	nb. of taxa (p)	5,635	595	329	100	59	36	18
	test statistic	13.5	8.3	9.2	8.2	7.2	5.6	5.3
	p-value	0.1070	0.5938	0.3017	0.1839	0.2046	0.3602	0.2429
Smoking	nb. of taxa (p)	7,793	595	278	84	51	31	15
	test statistic	19.5	23.8	19.0	14.2	12.7	13.5	14.3
	p-value	0.0048	0.0004	0.0019	0.0094	0.0108	0.0061	0.0016

Table F. Sensitivity analysis - Compositional equivalence test. Test statistic for high-dimensional data and one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

	Kingdom	Phylum	Class	Order	Family	Genus	Species	p-value _{adj}
Genus p = 142	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-002	NA	0.0129 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-003	NA	0.0129 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-005	NA	0.0252 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-UCG-010	NA	0.0949 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus-1	NA	0.0725 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae-NK4A214-group	NA	0.0376 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Christensenellaceae-R-7-group	NA	0.0376 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospira	NA	0.0129 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae-UCG-001	NA	0.0376 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae-UCG-010	NA	0.1884 (+)
	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae-NK4A136-group	NA	0.0376 (+)
	Bacteria	Tenericutes	Mollicutes	NB1-n	NA	NA	NA	0.0129 (+)

Table G. Sensitivity analysis - Smoking prevention experiment results. Differentially abundant taxa and adjusted Fisher p-values for 10,000 iterations at 5% prevalence filtering. Selected adjusted p-values ≤ 0.2 (sign of abundance difference: $y(1) - y(0)$).

Acknowledgements

I was engaged in researching and writing this thesis since 2018. The aim of my work is to present causal inference applications with the Rubin Causal Model framework in cases where only observational data is available. I became interested in causal inference methods after following Donald Rubin and Tirthankar Dasgupta's Design of Experiments class in Fall 2016. This is where I met Marie-Abèle Bind, while she was starting up her Lab at Harvard University's Department of Statistics. She gave me the opportunity to work with her on projects using causal inference methods to discover environmental effects and health outcomes. I would like to thank Marie-Abèle for her availability, support, and advice, which helped me find solutions to move forward. She is my first mentor and I will always keep her writing and critical thinking tips in mind. Problem-solving became much more fun since I have met Marie-Abèle! Thanks to her, I had the opportunity to get out of many comfort zones.

A special thought to the devoted students whom I had the pleasure to teach, in Spring 2020, during my last months at Harvard. The STAT140 - Design of Experiments class is the one that inspired me to do research in the first place. Your eagerness to learn has made every minute of extra preparation worthwhile.

I would also like to thank Annette Peters, Professor at the Medical Faculty of the LMU, for accepting to be my doctoral supervisor and for providing me with very valuable epidemiological guidance and feedback. I additionally am indebted to Christian Müller, Professor at the Department of Statistics of the LMU, for his interesting feedback during our weekly meetings, which helped push the last chapter of my thesis in a direction that is interesting for microbiome researchers.

An enormous thank you to my parents, for supporting my education, for enabling opportunities that have led me here, and for all the love you give me in your own way. My deepest gratitude to my husband, Edzi, who tried very hard to motivate me during the last and most difficult steps of finishing this thesis.

