Environmental Pollution 230 (2017) 730–740

Contents lists available at ScienceDirect

Environmental Pollution

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

Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among U.S. children[☆]

Jeanette A. Stingone [a], Om P. Pandey [b], Luz Claudio [a], Gaurav Pandey [b, c, *]

[a] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, USA
[b] Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, USA
[c] Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, USA

ARTICLE INFO

Article history:
Received 29 March 2017
Received in revised form
7 July 2017
Accepted 7 July 2017
Available online 18 July 2017

Keywords:
Multiple exposures
Mixtures
Machine learning
Neurodevelopment

ABSTRACT

Data-driven machine learning methods present an opportunity to simultaneously assess the impact of multiple air pollutants on health outcomes. The goal of this study was to apply a two-stage, data-driven approach to identify associations between air pollutant exposure profiles and children's cognitive skills. Data from 6900 children enrolled in the Early Childhood Longitudinal Study, Birth Cohort, a national study of children born in 2001 and followed through kindergarten, were linked to estimated concentrations of 104 ambient air toxics in the 2002 National Air Toxics Assessment using ZIP code of residence at age 9 months. In the first-stage, 100 regression trees were learned to identify ambient air pollutant exposure profiles most closely associated with scores on a standardized mathematics test administered to children in kindergarten. In the second-stage, the exposure profiles frequently predicting lower math scores were included within linear regression models and adjusted for confounders in order to estimate the magnitude of their effect on math scores. This approach was applied to the full population, and then to the populations living in urban and highly-populated urban areas. Our first-stage results in the full population suggested children with low trichloroethylene exposure had significantly lower math scores. This association was not observed for children living in urban communities, suggesting that confounding related to urbanicity needs to be considered within the first-stage. When restricting our analysis to populations living in urban and highly-populated urban areas, high isophorone levels were found to predict lower math scores. Within adjusted regression models of children in highly-populated urban areas, the estimated effect of higher isophorone exposure on math scores was −1.19 points (95% CI −1.94, −0.44). Similar results were observed for the overall population of urban children. This data-driven, two-stage approach can be applied to other populations, exposures and outcomes to generate hypotheses within high-dimensional exposure data.

1. Introduction

There is growing evidence that early-life exposure to ambient air pollution may affect neurodevelopment in children. Epidemiologic studies have shown that prenatal and/or early-life exposures to ambient air pollutants are associated with measures of neurodevelopment and behavior in infants and young children (Edwards et al., 2010; Freire et al., 2010; Guxens et al., 2012; Lin et al., 2014; Perera et al., 2006, 2009), autism diagnoses (Becerra et al., 2013; Jung et al., 2013; Kalkbrenner et al., 2010; Roberts et al., 2013; Volk et al., 2013, 2014; Windham et al., 2006) and attention-deficit/hyperactivity disorder (Newman et al., 2013). There is also evidence that air pollutants contribute to deficits in neurodevelopment that persist into later childhood (Suglia et al., 2008), affecting cognitive outcomes such as academic achievement. Although ambient air is a complex mixture of multiple pollutants, most of this previous research has focused on associations between individual pollutants and children's cognitive health (Becerra et al., 2013; Edwards et al., 2010; Freire et al., 2010;

[☆] This paper has been recommended for acceptance by David Carpenter.

[*] Corresponding author. One Gustave L. Levy Place, Box 1498, New York, NY 10029, USA.
E-mail address: gaurav.pandey@mssm.edu (G. Pandey).

http://dx.doi.org/10.1016/j.envpol.2017.07.023

Guxens et al., 2012; Jung et al., 2013; Lin et al., 2014; Newman et al., 2013; Perera et al., 2006, 2009; Roberts et al., 2013; Suglia et al., 2008; Volk et al., 2013, 2014; Windham et al., 2006). Environmental epidemiology is now transitioning from single-pollutant approaches to more holistic investigations of the exposome and environment's collective effect on health. The recent availability of datasets containing exposure estimates for multiple air pollutants, population demographics and health outcomes on large cohorts of children provides an opportunity to leverage methods for “big data” to advance environmental epidemiology (Bellazzi, 2014).

In a 2014 review, Oakes et al. identified fifty-seven distinct studies that focused on developing multi-pollutant metrics of exposure for a variety of outcomes (Oakes et al., 2014a). The authors noted a lack of consensus on which multi-pollutant metrics were recommended for a given research question. They identified that a key limitation is that most metrics assume pure additivity of effects with no potential for synergistic or antagonistic interactions. This can be a major limitation, since pollutants vary spatially and can combine with each other to create distinct mixtures that may have different effects on exposed populations than the individual pollutants. Identification of these spatially-varying exposure profiles may allow researchers to pinpoint affected communities and target more in-depth research into sources and potential health effects. For example, Coker et al. used Bayesian profile regression to identify exposure profiles associated with adverse birth outcomes in Los Angeles (Coker et al., 2016). That study examined only three pollutants in conjunction with contextual neighborhood factors that could simultaneously impact birth outcomes.

Machine learning (ML) methods can be used to identify the exposures relevant to health outcomes of interest within high-dimensional exposure data, as well as the potential interactions between those exposures (Patel, 2017). A recent application of ML methods, specifically classification and regression tree (CaRT) (Lemon et al., 2003), in air pollution epidemiology by Gass et al. examined the relationship of a small number of pollutants to asthma emergency department (ED) visits (Gass et al., 2014). In that study, the typical CaRT objective of predicting the dependent variable (here, use of the ED) was replaced by identifying statistically significant combinations of (discrete) pollutant levels that best capture the risk of asthma ED visits as compared to referent levels of the pollutants. Although a promising step forward, this work confounds the goals of prediction using CaRT methods and estimation of effect sizes of the contributing pollutant combinations. It may be more appropriate to use CaRT methods as an initial screening tool to identify combinations of interest and then use a second analytic method to estimate the effect size, as suggested by Sun et al. in their recent review (Sun et al., 2013). Using CaRT as a first-stage method allows for the examination of continuous exposure variables, as opposed to arbitrary discretization of the exposures. Additionally, this method can provide a more stable picture of the association between a pollutant profile and the outcome of interest than a single tree by learning multiple trees and then examining the occurrence frequency of the pollutant profile within slightly different samples from the study population. The pollutant profiles identified in the first-stage can then be investigated in more depth in the second-stage by using well-established epidemiologic methods to control for confounding, assess effect measure modification and investigate various exposure contrasts.

The goal of our study was to apply a data-driven approach to identify early-life exposure profiles associated with measures of cognitive skills and school readiness in a nationally-representative cohort of 6900 U.S. children (Najarian et al., 2010). This two-stage approach incorporates machine learning into environmental health research by first using CaRT methods to identify pollutant profiles associated with test scores. Then, epidemiologic methods for effect estimation and assessment of interaction were used to quantify the magnitude of the combined effect of these pollutant profiles on the children's math scores. We applied this approach in combination with stratification based on urbanicity levels, which were expected to confound the relationship between air pollution exposure and early cognitive skills.

2. Materials and methods

2.1. Study population

Conducted by the National Center of Education Statistics, the Early Childhood Longitudinal Study, Birth cohort (ECLS-B) is a longitudinal study of a nationally representative, random selection of children born in 2001 and followed from the age of 9 months through kindergarten entry (Najarian et al., 2010). Women and children were recruited from birth certificate data and contacted for study visits at 9 months, 2 years, 4 years, and during kindergarten. At each visit, children participated in neurodevelopmental assessment activities and mothers participated in interviews. At later study points, childcare providers and teachers also participated in interviews. All sample sizes mentioned subsequently in this article are rounded to the nearest 50 to comply with ECLS-B privacy guidelines. Approximately 74% of eligible women and children ($N = 10{,}700$) agreed to participate at study entry. Of these children, approximately 83% completed the preschool and kindergarten assessments at 4 and 5 years of age ($N = 8900$). For this study, children were limited to singleton births, whose mother provided a residential address at study entry and who completed the study assessments during Kindergarten, resulting in a cohort of 6900 children.

2.2. Outcome assessment: mathematics standardized tests

At the kindergarten study visit, each child completed a variety of standardized tests aimed at assessing their basic math and verbal skills, as appropriate for school-entry. Because math scores may be less prone to confounding from language spoken in the child's home (Roberts and Bryant, 2011), our study utilized math scores as the primary outcome. The 58-item mathematics assessment was derived from standardized instruments, including the Test of Early Mathematics Ability (TEMA-3) and mathematics assessments from other NCES childhood studies. The concepts covered in the assessment included number sense, properties, operations, measurement, geometry, spatial sense, data analysis, statistics, probability, patterns, algebra, and functions (Najarian et al., 2010). An adaptive two-stage design was used to adjust the test-difficulty based on the number of correct responses during the initial stage of the assessment. As the goal of our study was to assess the association between math scores and exposure to ambient air toxics, the raw scale score was used in all analyses, as suggested by NCES analytic guidelines.

2.3. Exposure assessment: estimated concentrations of ambient air toxics

Exposure to air toxics was assigned using data derived from the U.S. Environmental Protection Agency's National Air Toxics Assessment (NATA) (EPA, 2013). Air toxics, also known as hazardous air pollutants, are listed in the Clean Air Act and thought to be