Who Is Most Vulnerable? A Causal Machine Learning Approach for Environmental Justice*

Falco J. Bargagli Stoffi

Harvard University, Boston, MA, USA

Background. The U.S. Environmental Protection Agency (EPA) has set the goal to achieve environmental justice by addressing the disproportionate vulnerabilities to adverse human health effects resulting from exposure to air pollution. According to the EPA, environmental justice means that ""no group of people should bear a disproportionate burden of environmental harms and risks." In pursuit of this goal, the EPA has called for scientific studies that would shed light on demographic-specific information and the disproportionate health impacts of air pollution. This work aims to answer the EPA call by providing nationwide data-driven evidence regarding the most vulnerable subgroups to exposure to air pollution.

Methods. We developed a new method in causal inference and machine learning to identify which subgroups of the Medicare population are most vulnerable or resilient to long-term exposure to fine particulate matter ($PM_{2.5}$) on mortality. A key feature of this approach, which we call Causal Rule Ensemble (CRE), is that it enables the interpretable, data-driven discovery of vulnerability/resilience in epidemiology studies. We acquired and integrated data on 35,331,290 Medicare beneficiaries across the entire United States for the period 2010-2016. We considered a binary exposure, indicating whether each individual has been exposed to $PM_{2.5}$ greater than the current NAAQS of 12 micrograms per cubic meter or not. We linked exposure to two-year annual $PM_{2.5}$ during 2010-2011 at the zip code level to mortality during the five years 2012-2016 and several potential confounders. Finally, we studied the heterogeneity in the causal effects in the four U.S. geographic regions: Northeast, Midwest, West, and South.

Results. The data-driven causal machine learning approach revealed subgroups at even higher risk with respect to the population risk. We found evidence of increased risk for rural (low-population density) communities in the Midwest, Northeast, and South; for the black community in the South; for low-income and less educated communities in the Northeast. However, some minority subgroups showed potential resiliency to air pollution exposure. While surprising, this decrease in mortality when exposed to higher levels of air pollution has already been documented in the literature as evidence of survival bias.

Conclusions. We propose a new methodology for discovering vulnerability or resilience to air pollution exposure. Our causal machine learning model identifies the key factors in the individual characteristics that explain different degrees of vulnerability or resilience to air pollution. This work opens a new avenue of research at the intersection of causal inference, artificial intelligence, and epidemiology in service of environmental justice. Open software is made publicly available via the CRE R package on CRAN (https://cran.r-project.org/web/packages/CRE/index.html).

*Study not funded by HEI