Big, Messy Data – Getting to Clean Air

By: Orna Feldman

Vision is the art of seeing what is invisible to others.
-Jonathan Swift

Trying to get a grip on the imperceptible is the large-frame focus for a small team of Harvard T.H. Chan School of Public Health researchers. Air pollution—carbon dioxide, particulate matter, toxic organic compounds and the other unhealthy substances we routinely breathe— is their métier. Their method, statistical, leverages an innovative mathematical model and array of skills spurring progress other groups haven't even attempted to achieve.

A key leader in the effort is Cory Zigler, PhD, assistant professor in the Department of Biostatistics. With his co-leader and mentor Francesca Dominici, a biostatistical force in air pollution studies and the School's senior associate dean for research, Zigler brings to the problem unique expertise: Comparative effectiveness research, know-how in extracting policy-relevant information from data, and perhaps most salient, fresh perspective. His multi-pronged skill set garnered a three-year, $1 million grant from the independent research organization Health Effects Institute while he was a post-doctoral research fellow, an impressive achievement so early in one's career ("A combination of luck and timing," claims Zigler). The Zigler/ Dominici duo, backed by top-level scientists from disparate disciplines, is driving a big-data study whose results will critically impact the future of anyone in the United States who breathes.

Casting the problem epidemiologically, Dominici quotes her grandmother: "Air quality isn't like choosing to eat a donut. Everybody breathes it. Air impacts everything we do, is expensive to clean and affects climate change now and for the next 100 years."

The financial advantage of reducing air pollution is gargantuan. The estimated cost of the benefits accrued by all EPA air quality regulations between 2002 and 2012 ranges between $109.4 billion and $629.1 billion. In a Science commentary, Dominici and her co-authors reported that reducing one single pollutant–particulate matter–accounted for one-third to one-half the total monetized benefits of all air pollution federal regulations in that time period. By 2020, the costs to improve air quality are predicted to skyrocket to $2 trillion. A key culprit of air pollution is power plants, which account for one-third of all greenhouse gas emissions.

In an effort to enhance air quality, the Obama administration recently called for a 30% reduction in power plant-generated pollution by 2030. New air pollution regulations, among the costliest–and most beneficial–of all federal regulations, could eventually shutter hundreds of coal-fired plants. Delaying action to reduce CO2 emissions could cost $150 billion yearly, according to a recent White House report. Correct estimates of these regulations' health impacts will affect a cascade of regulatory decisions to come, making continued research and analysis critical in resolving uncertainties.

Timing

In the forefront of identifying cost-effective solutions that maximally benefit public health are Zigler, Dominici and their collaborators. Zigler's heft in the field comes from a confluence of factors. One is timing. From the early 1970s, air pollution research was propelled by fundamental questions such as, "Can we prove air pollution is bad for people? If so, how bad? What's the relationship between air pollution exposure and heart attacks, strokes, asthma and other diseases?" The state of air pollution research was akin to that of smoking research in the 1950s: The aim was collecting evidence that proved harm.

In the last few years, epidemiologists–as well as economists, statisticians, political scientists and other academicians–began to re-frame their questions. The focus began to shift to intervention: "What works best?" This new emphasis on quantifying responsi-

bility, consequential epidemiology as Zigler refers to it, relies on causal inference (CI), a statistical perspective that tries to tease out causal effects linking specific actions, rather than associations, from a plethora of data.

The CI approach takes into account a host of downstream outcomes, folding in a multi- layered complexity to air pollution science. In their power plant research Zigler and Dominici grapple with a constellation of factors: weather; behavior patterns; energy use patterns; atmospheric chemistry producing pollution; the proximity of two power plants; the distance a person lives from a power plant; the installation of power plant scrubbers (devices controlling gas emissions); 50 million annual Medicare billing claims associated with hospitalization and mortality, and limitless other variables.

Fresh Perspective

While masterful in biostatistics and CI, Zigler did not have a background in environmental health; his expertise lay in applying biostatistics to clinical studies of HIV/AIDS, substance abuse and maxillofacial surgery. Shifting his CI lens to epidemiology put him against the grain of traditional air pollution researchers – "I was a bit of a fish out of water, and I kept advancing it," he says. His CI expertise has become formidably effective in re-orienting air pollution research. By twisting the questions and considering comparable alternatives, he and his team amass data towards one aim: Identifying which air quality interventions targeted to power plants yield the best health outcomes, a crucial step in establishing cost-effectiveness.

The available data are massive. But they come from a patchwork of incompatible, broadly sourced, and poorly utilized databases collected from diverse government and public health agencies. In short, the data aren't integrated. Coupled with the use of limited analytic tools, these shortcomings critically hinder big data's ability to improve people's health.

"We know with a high degree of accuracy the emissions of different pollutants for all power plants in the US. We know whether a power plant facility has taken any action to reduce its emissions and what type of action. But no one has put together the data of what's been done, the downstream effects of implementing one intervention over another, the costs, the results with the biggest bang," says Dominici. "No one has studied whether any of these actions have actually helped."

We have information about the pollutants emitted by all US power plants, as well as the steps taken by each to reduce emissions, but no data exists on the effects of implementing one intervention over another.

The EPA, intent on knowing the effects of regulatory interventions to support policy decisions ,will be awarding $30 million to answering the question, Zigler and Dominici are waiting to hear the status of their proposal combining CI methodology, atmospheric modeling, new analytic tools and national datasets. Their team is collaborating with the EPA/Harvard University Center for Ambient Particle Health Effects, headed by world-wide expert in atmospheric science and exposure assessment Petros Koutrakis, Professor of Environmental Health at the Harvard T.H. Chan School. Citing the Zigler/Dominici advantage, he says "they address an important science and policy issue systematically, in a way other [researchers] haven't done." Their research results will offer new evidence with enhanced specificity and unprecedented accuracy, emphasize interventions, and create a national analytic system for population health research.

Expertise Across the Board

A number of other factors favorably position Zigler in air pollution clean-up efforts. Many seasoned CI researchers don't have any familiarity with the complex federal regulatory system. "I'm not a regulatory expert," he says, but the statistical training he gained in the clinical setting predisposed him to "see a disconnect between what regulators were asking– 'Is a particular air quality action effective in reducing air pollution and saving lives?' –and the questions epidemiologists have been answering–'Does a hypothetical decrease in air pollution associate with an increase in life expectancy?' I was convinced the regulatory system didn't want only information about how bad pollution is. What the regulatory system is looking for are evidence-based evaluations of their air quality control actions. The answers are essential to learning what the system could actually do about it."

Adding fuel to the team's anti-pollution power is a group of top-notch researchers drawn from many quarters. Their ranks include academicians from wide-ranging disciplines, from levels ranging from PhD to accolade-heavy senior scientists, including Professor Gary King, director of Harvard University's Institute for Quantitative Social Science, and Dr. Jonathan Samet of the University of Southern California. Along with the other assembled researchers–Koutrakis, epidemiologist Joel Schwartz and data scientist Christine Choirat among them–they introduce slightly different questions and answers, collectively expanding the study's approach. "I'm interested in advancing methods for causal inference that address public health questions at the interface at of all these disciplines." Says Zigler. Adds Dominici: Zigler's forte is "speaking many languages, integrating knowledge across many domains and translating" one discipline to another.

For his part, Zigler views himself as "standing on the shoulders of giants," with Dominici being a key shoulder. "Her mentorship has been invaluable. Not only does she have the chops in air pollution and biostatistics, she has seats at all the regulatory and funding tables and is extremely talented in focusing on gaps in scientific knowledge. I've benefitted from her more than I ever imagined."

Benefit Writ Large

The larger social benefit promises to be "paradigm-shifting," says Dominici, who sees their work as the public health analogue to big-data's ground-breaking impact on the biological sciences. "A national data system informed by new mathematical models and analyses is a radical concept in population health science," she says. Revealing the evidence, and slicing and dicing it with new models, allows previously imperceptible patterns to become visible, leading to more reliable problem solving. "Our work offers a totally new dimension to complement the science supporting air quality policies," adds Zigler. "We can't get rid of air pollution entirely. But we can learn which actions–real actions, not hypotheticals–are most effective in protecting people from breathing dirty air."