SUPPLEMENTARY MATERIAL

Hypothesis: Analyzing Individual-Level Secondary Data with Instrumental Variable Methods is Useful for Studying the Effects of Air Pollution on Dementia

KELLY C. BISHOP

Arizona State University, Department of Economics

SEHBA HUSAIN-KRAUTTER

Mount Sinai School of Medicine

JONATHAN D. KETCHAM

Arizona State University, Department of Economics

NICOLAI V. KUMINOFF

Arizona State University, Department of Economics,
and National Bureau of Economic Research

CORBETT SCHIMMING

Mount Sinai School of Medicine,
and James J. Peters Veterans Affairs Medical Center

1. Claims-based definitions of Alzheimer’s disease and dementia

The World Health Organization’s 10th revision of the International Statistical Classification of Diseases and Related Health Problems defines dementia (codes F00-F03) as “a syndrome due to disease of the brain, usually of a chronic or progressive nature, in which there is disturbance of multiple higher cortical functions, including memory, thinking, orientation, comprehension, calculation, learning capacity, language and judgement. Consciousness is not clouded. The impairments of cognitive function are commonly accompanied, and occasionally preceded, by deterioration in emotional control, social behavior, or motivation. This syndrome occurs in Alzheimer disease, in cerebrovascular disease, and in other conditions primarily or secondarily affecting the brain” (World Health Organization 2011).

The ICD-10 defines Alzheimer’s disease (G30) as “A degenerative disease of the brain characterized by the insidious onset of dementia. Impairment of memory, judgment, attention span, and problem solving skills are followed by severe apraxias and a global loss of cognitive abilities. The condition primarily occurs after age 60, and is marked pathologically by severe cortical atrophy and the triad of senile plaques; neurofibrillary tangles; and neuropil threads” (World Health Organization 2011).

CMS insurance claims data include these diagnostic codes, and CMS itself uses them to track the evolution of individuals’ health in their Chronic Condition Warehouse (CCW). The claims data and the CCW data are available to researchers. The CCW defines Alzheimer’s disease by looking for the presence of any of these codes in the prior three years on a claim from most health care providers: DX G30.0 G30.1, G30.8 G30.9, and prior to October 2015 the ICD-9 code DX 331.0. Researchers with access to the individual Medicare claims could adapt this definition as they see appropriate.

The CMS identifies Alzheimer’s disease and related dementias using a broader set of diagnostic codes. Currently these are ICD-10s DX F01.50, F01.51, F02.80, F02.81, F03.90, F03.91, F04, G13.2, G13.8, F05, F06.1, F06.8, G30.0, G30.1, G30.8, G30.9, G31.1, G31.2, G31.01, G31.09, G91.4, G94, R41.81, R54. For data prior to October, 2015 they use ICD-9s DX 331.0, 331.11, 331.19, 331.2, 331.7, 290.0, 290.10, 290.11, 290.12, 290.13, 290.20, 290.21, 290.3, 290.40, 290.41, 290.42, 290.43, 294.0, 294.10, 294.11, 294.20, 294.21, 294.8, 797.

MCI is not currently tracked within the CCW data, but researchers can use the presence of diagnostic code G31.84 on the claims to identify the presence and timing of a diagnosis for MCI.

2. Monte Carlo Simulation

The article explains how IV estimation can overcome potential confounding from measurement errors and unmeasured variables. In addition to these advantages, however, IV estimation also reduces statistical precision and adds a "finite sample bias" that diminishes as the sample size grows. These properties make IV estimators well-suited for secondary data sets in which researchers can leverage large samples to overcome concerns about statistical imprecision and finite sample bias.

We present a Monte Carlo experiment to show how IV estimation can mitigate confounding and how its performance varies with sample size. In this context, Monte Carlo represents a statistical experiment to demonstrate the advantages of IV estimation over non-causal methods in the presence of error in the measurement of pollution, error in the measurement of dementia, and unmeasured risk factors for dementia and different degrees of finite sample bias. Because such errors cannot be observed by researchers in existing data, we use Monte Carlo methods to generate data in which the errors and other data characteristics are created by us to assess the performance of different estimators under different conditions.

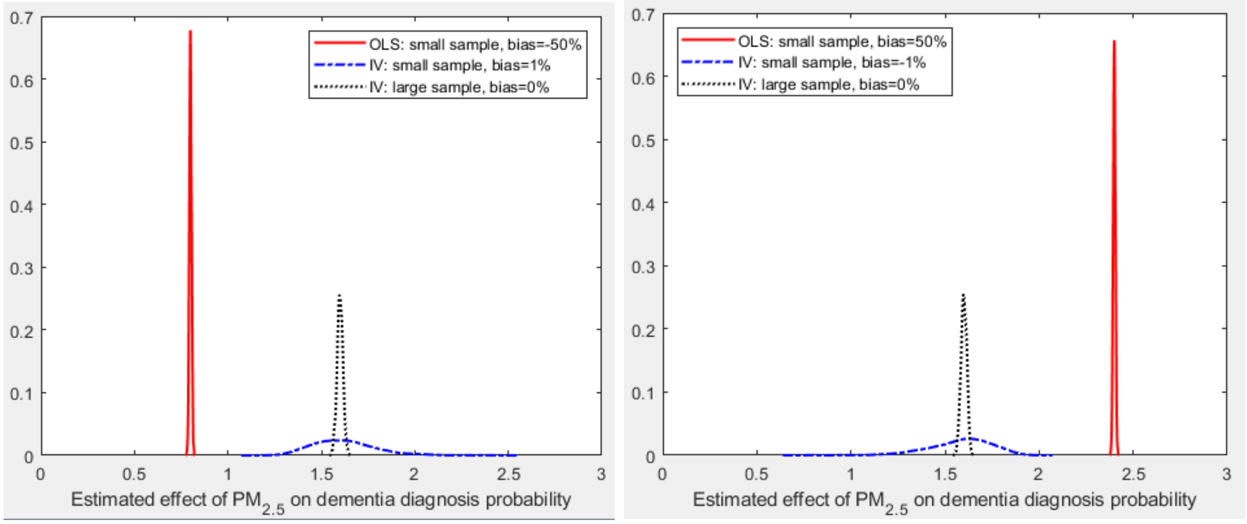
We use a random number generator to generate data with joint distributions for decadal average $PM_{2.5}$ concentration (Mean=10, Variance=1), an IV (Mean=0.3, Variance=1), and a composite model error (Mean=0, Variance=1) that represents the net effect of measurement errors and unmeasured risk factors. The intercept and slope of a linear equation that we use to represent the true but unknown relationship between $PM_{2.5}$ inhalation and dementia onset are chosen so that 20% of the population has dementia and a 1 $\mu g/m^3$ increase in decadal $PM_{2.5}$ concentration increases the risk of dementia onset by 1.6 percentage points. We define the IV to be negatively correlated with actual $PM_{2.5}$ exposure (ρ=-0.065) and uncorrelated with the model error. Then we randomly draw 1,000 samples from these data to evaluate how different research designs yield different estimates of $PM_{2.5}$'s effect on the probability of an incidental dementia diagnosis: (1) a benchmark small-sample ordinary least squares (OLS) regression where each sample has size N=10,000; (2) a small-sample IV regression with each sample has size N=10,000, and (3) a large-sample IV regression with N=1,000,000. [1]

Figure A1 shows the distributions of the 1,000 estimates for $PM_{2.5}$'s effect on dementia for each of the three estimators in two different versions of the Monte Carlo simulation. In case 1 we set the model error to be negatively correlated with $PM_{2.5}$ (ρ=-0.8) and in case 2 we set the model error to be positively correlated with $PM_{2.5}$ (ρ=0.8). In case 1, the small-sample OLS estimator that ignores confounding problems has a high degree of statistical precision but a large downward bias, equivalent to 50% of the true 1.6 percentage point effect. The small-sample IV estimator eliminates all but 1% of this bias, but it has a relatively large variance. Hence, when the IV estimator is applied to a small sample, the point estimate may be above or below is biased toward the OLS estimate, and the confidence interval will be wide compared to OLS. These problems with the IV estimator diminish with the larger sample size. With N=1,000,000, the IV estimator has a bias of less than one thousandth of a percentage point and the distribution of 1,000 estimates has a variance that is

[1] While our example explicitly shows how the IV estimator works in linear regression models, it is also representative of how IV estimators work in nonlinear models such as a probit.

similar to the small-sample OLS estimator. The IV estimator's performance is essentially the same in case 2 when the OLS estimator that ignores confounding has an upward bias of 50%.

Figure A1. Results from a Monte Carlo experiment under different conditions comparing estimates from ordinary least squares and instrumental variables estimators



<u>Case 1</u>: $\boldsymbol{corr(PM_{2.5}, model\ error) = -0.8}$ **<u>Case 2</u>:** $\boldsymbol{corr(PM_{2.5}, model\ error) = 0.8}$

<u>Note:</u> Each distribution represents 1,000 estimates of the effect of $PM_{2.5}$ on the probability of an incidental dementia diagnosis from each of the three models described in the legend and text.