COMMENT BY

DOUGLAS O. STAIGER\(^1\) Over the last four decades a large literature has documented substantial variation in health care spending, utilization, and quality across regions, hospitals, and physicians (IOM 2013, Skinner 2012). Many analysts, and the Dartmouth group in particular, argue that these differences reflect persistent productivity differences, and are not the result of random fluctuation or unmeasured differences in patient health status or preferences.

My figure 1 provides a typical example of the kind of evidence that has been produced. It plots hospital-level data on mortality and spending for Medicare patients admitted to the hospital with a heart attack (acute myocardial infarction, or AMI) during 2007–09. As is common in this literature, the hospital-level estimates were adjusted using standard methods to control for observed differences in patient risk and hospital prices, and hospitals with fewer than 200 AMI patients were not plotted to minimize random fluctuations (Skinner and others 2013). Nevertheless, there remains substantial variation across hospitals: average 1-year Medicare expenses range from $30,000 to $60,000, while 1-year mortality rates range from 20 percent to 45 percent. Compared to the most productive hospitals (indicated by the scatterplot points in the lower left of the figure), patient mortality and expenditures are roughly a third higher in the least productive hospitals (upper right).

The key features of my figure 1—twofold variation in cost and quality, which are largely unrelated to each other, implying large variation in productivity—are not unique to this example. Similar results have been

---

1. I thank Amitabh Chandra, Jonathan Skinner, and James Feyrer for helpful comments and advice. I am co-founder of, have an equity interest in, and have consulted for ArborMetrix, a company that sells performance measurement systems and consulting services to insurers and hospitals.
found across different levels of aggregation (county, state, hospital, physician) using a wide variety of data sets and patient populations. In fact, persistent productivity differences are not unique to health care. A long literature has documented similarly large productivity differences across plants and firms (Syverson 2011), teachers and schools (Jackson, Rockoff, and Staiger 2014) and countries (Klenow and Rodríguez-Clare 1997). In health care, as in these other settings, there is a growing consensus that these differences are in large part causal and capture how average costs and outcomes would change if a patient were to be treated in a different region or by a different provider. Much of the recent literature has moved on to understanding why these productivity differences exist and developing strategies that would encourage less efficient providers and regions to move, in Peter Orszag’s

Figure 1. Spending and Mortality for AMI Admissions, by Hospital, 2007–09

Source: Author’s calculations, using data from Skinner and others (2013).

a. The figure plots risk-adjusted mortality rates against risk- and price-adjusted 1-year Medicare expenditures (dollars) for each hospital, based on Medicare beneficiaries age 65 or over admitted for an acute myocardial infarction (AMI) from 2007 to 2009. Only hospitals with at least 200 Medicare AMI admissions are plotted.

b. Risk and price adjustment were done using standard patient-level methods; risk adjustment did not control for Medicare’s Hierarchical Condition Categories (HCCs). See Skinner and others (2013).
words, toward “the proven and successful practices adopted by lower-cost areas and hospitals” (Orszag 2009).

In the paper under discussion here, Louise Sheiner takes a step back to challenge this consensus. She argues that geographic variation in health care spending reflects unmeasured population health needs and socioeconomic characteristics rather than causal productivity differences. Her key piece of evidence, illustrated in her figure 1, is that there is a strong correlation between average Medicare spending in a state and proxies for health needs in the state such as the share of the state that is diabetic. This correlation, she argues, is *prima facie* evidence that geographic variation is driven primarily by differences in population health needs, not differences in productivity. After she controls for the share of the state that is diabetic and a few other state-level measures of socioeconomic status, little unexplained variation remains. More broadly, Sheiner argues that it is difficult to learn anything from geographic variation, because high-spending states differ in multiple dimensions from low-spending states.

It is not clear what we learn from the correlations documented by Sheiner. Earlier work has found similar correlations of population health and demographic measures with health care spending and quality, but the interpretation of these correlations has varied. For example, Richard Cooper (2009) documents a strong correlation between regional spending and poverty and, like Sheiner, has argued that regional variation reflects health needs rather than productivity differences. Alternatively, a number of papers have found higher mortality and adverse event rates in hospitals serving minority populations, and interpreted this as evidence that true quality of care is lower in minority-serving hospitals (Morales and others 2005, Skinner and others 2005, Lake and others 2015). Amitabh Chandra and Staiger (2007) find that areas with the types of heart attack patients most appropriate for aggressive treatment are more likely to provide aggressive treatment to all patients, but argue that this is the result of true productivity differences in regions that specialize in providing aggressive treatment. Jonathan Skinner and Staiger (2007) also find strong correlations at the state level between health care spending and mortality (for Medicare AMI patients) and a wide variety of demographic measures and proxies for innovativeness in the state, but they interpret the data as suggesting that true productivity differences are related to state-level barriers to adoption of efficient technologies (see also Skinner and Staiger 2015).

Are these correlations driven by unmeasured health differences across areas, as Sheiner and Cooper both argue, or is there simply a correlation between true productivity differences and state population characteristics,
as much of the recent literature argues? A simple empirical framework that relates the standard patient-level estimation method to Sheiner’s state-level estimation method helps to clarify why the state-level correlations documented by Sheiner cannot answer this question.

The standard risk-adjustment method runs a patient-level fixed-effect regression of the form

\[ Y_{ij} = \beta_0 + \alpha_j + \epsilon_{ij}, \]

where \( Y_{ij} \) is a patient outcome (such as spending or mortality) for patient \( i \) treated in region (or hospital) \( j \), and \( X_{ij} \) is a vector of patient-level controls such as age, gender, race, pre-existing medical conditions and, in some data sets, detailed clinical characteristics. Estimates of the fixed effects \( \hat{\alpha}_j \) are the risk-adjusted measures of patient spending or mortality in each region or hospital (such as what was plotted in my figure 1). These measures may be associated with average characteristics of patients \( X_j \) or other characteristics \( Z_j \) of the region:

\[ \hat{\alpha}_j = \bar{X}_j \beta_1 + Z_j \beta_2 + \hat{\omega}_j. \]

Averaging my equation 1 at the state level and substituting for \( \hat{\alpha}_j \) using my equation 2 yields Sheiner’s between-state estimator:

\[ \bar{Y}_j = \bar{X}_j (\hat{\beta}_1 + \hat{\alpha}_j) + Z_j \hat{\beta}_2 + \hat{\omega}_j. \]

Sheiner argues that the residual from this regression \( \hat{\omega}_j \) is a better estimate of the causal variation in productivity across states than the residual from my equation 1 \( \hat{\alpha}_j \). Thus, the key to understanding the difference between the standard approach and Sheiner’s approach is my equation 2: Sheiner believes that the correlation between \( \hat{\alpha}_j \) and the regional measures \( (\bar{X}_j, Z_j) \) is due to health attributes that were omitted from my equation 1 and removes it, whereas the standard approach treats \( \hat{\alpha}_j \) as true productivity estimates and tries to understand why productivity is correlated with the regional measures. Estimates of my equation 2 (or equivalently, Sheiner’s equation 3) provide no evidence to decide which interpretation is correct. Therefore, the state-level correlations documented by Sheiner cannot answer the key question of whether regional and provider variation is informative about productivity.

Other evidence, however, strongly suggests that regional and provider-level variation is not driven by omitted health attributes. First, regional and hospital-level estimates that control for more detailed clinical information
at the patient level are highly correlated with standard estimates, and they continue to be correlated with population characteristics such as income and race (Dimick and Birkmeyer 2008, Dimick and others 2010, McClellan and Staiger 2000, Morales and others 2005, Skinner and others 2005, Sutherland, Fisher, and Skinner 2009). Second, there is substantial variation across regions and hospitals in measures for which health attributes are not relevant, such as aspirin and beta blocker use in the hospital following a heart attack, and these measures are also strongly correlated with population characteristics such as income and race (Baicker and Chandra 2004, Skinner and Staiger 2007). Third, there is substantial variation across regions in how physicians say they would treat patients in standardized vignettes (holding patient attributes constant), and this variation is strongly correlated with actual practice patterns (Cutler and others 2013). Finally, two seminal papers from the Dartmouth group (Fisher and others 2003a and 2003b) focus on regional differences in end-of-life spending because it “reflects the component of regional variation in Medicare spending that is unrelated to regional differences in illness” (p. 273). They find that health status is similar across regions with different end-of-life spending, but patients received roughly 60 percent more care in the higher-spending regions.

More importantly, there is growing evidence that regional and provider variation in spending and patient outcomes are related to causal estimates of productivity differences. First, there are large impacts on average costs when patients are randomized to medical groups (Doyle, Ewer, and Wagner 2010) or quasi-randomized to hospitals (Doyle 2011, Doyle and others 2015), and these differences are strongly predicted by standard risk-adjusted estimates. Second, Amy Finkelstein, Matthew Gentzkow, and Heidi Williams (2014) find that when Medicare patients move from one region to another, their Medicare spending and utilization change immediately by an amount that is roughly 50 to 60 percent of the difference in average spending and utilization between the regions. Consistent with Jason Sutherland, Elliott Fisher, and Jonathan Skinner (2009), they find that health characteristics can explain only about a third of the average differences in spending and utilization across regions. Finally, John Birkmeyer and others (2013) directly assesses surgical technical skill using video, and find that surgeon skill was strongly negatively associated with risk-adjusted patient complication rates and subsequent rates of reoperation, readmission, and emergency room admission.

All of this evidence suggests that standard patient-level risk-adjusted measures of spending, utilization, and quality are to a large extent capturing
causal productivity differences across regions and providers rather than omitted health attributes of the population. This would suggest that Sheiner is finding a correlation between true productivity differences and state population characteristics, and misinterpreting it as evidence of omitted health attributes.

This view of Sheiner’s results is consistent with a broader literature arguing that micro-founded estimates of productivity are valid, and that macro-level correlations of productivity with population characteristics are difficult to interpret. For example, in the field of education, so-called value-added estimates of teacher performance (based on student-level test score regressions similar to my equation 1) have been validated in a number of natural and randomized experiments (Bacher-Hicks, Kane, and Staiger 2014, Chetty, Friedman, and Rockoff 2014a, 2014b, Kane and Staiger 2008, Kane and others 2013), despite the fact that these estimates are strongly correlated with average student poverty. Macroeconomists have abandoned production function estimates based on cross-country variation (which, like Sheiner’s cross-state regressions, explain away most of the variation), and instead use estimates of total factor productivity based on micro-founded estimates of the returns to physical and human capital (Klenow and Rodríguez-Clare 1997).

The important question is not whether large and persistent productivity differences exist—the evidence is overwhelming that they do. The important question is why these productivity differences exist, why they are associated with a number of market and firm characteristics, and whether there are strategies that could improve productivity and welfare. As Sheiner notes, and as has become apparent in the macro literature on productivity, simple correlations between productivity and market characteristics are not likely to take us very far. Instead, we need models that can explain persistent productivity differences, empirical tests of these models, and evaluations of interventions designed to improve productivity. Along these lines, recent work on productivity spillovers (Chandra and Staiger 2007), technology diffusion (Skinner and Staiger 2015), competition (Bloom and others 2010, Chandra and others 2013), and management quality (Bloom and others 2014) are a promising step forward toward a better understanding of why these differences exist. Running cross-section regressions of state-level spending on state-level health measures, as Sheiner does, is a step backwards.

REFERENCES FOR THE STAIGER COMMENT
Baicker, Katherine, and Amitabh Chandra. 2004. “Medicare Spending, the Physician Workforce, and Beneficiaries’ Quality of Care.” *Health Affairs*, April.


COMMENTS and DISCUSSION


GENERAL DISCUSSION

Peter Orszag agreed with discussant David Cutler that even without emphasizing the role of regional variation one must conclude that there is a massive amount of variation in the practice norms in health care and that risk adjustments do not explain it. In fact, hospitals are hiring businesses to examine their practices and identify risk-adjusted clinical variation to reduce wasteful practices, something that would make no sense to do if all the treatment variation stemmed from the riskiness of the patients alone. This approach is becoming more prominent for two reasons. First, hospitals had little incentive to do it as long as they did not face capitated or other risk-based payment systems; however, the reality and, more importantly, the perception now is that the payment system is evolving away from fee-for-service payments. Second, digitization of medical records makes these analyses easier to do than in a paper-based system.

Orszag wondered how variations in cost relate to treatment quality. He cited a series of studies done by Joseph Doyle and his colleagues. Their initial paper suggested that higher-cost hospitals showed lower mortality rates one year after acute-care treatment, but that paper focused only on emergency room visits. When an updated version of the analysis looked