What Healthcare Teaches Us About Measuring Productivity in Higher Education

Douglas Staiger[1]

Dartmouth College and NBER

I. Parallel problems, parallel lessons

As discussed in the introduction to this volume, any study that attempts to measure productivity in higher education faces numerous challenges. Knowing these challenges and laden with institutional knowledge, higher education experts may be tempted to "go it alone" in studying productivity. They may even, feeling that the challenges are insurmountable, refrain from studying productivity in higher education at all. However, many of the same issues arise when studying productivity in the health care industry, and there is a rich history of researchers confronting and overcoming these issues. It would be wasteful not to distill the lessons learned in health care and suggest how they apply to higher education. Thus, in this chapter, I identify parallels between the health care and higher education sectors. I suggest lessons from health care that might translate to the study of productivity in higher education.

II. Measuring productivity in health care: a central example

To help make the discussion that follows concrete, especially for a higher education audience less familiar with health care, let us begin with a typical exercise in measuring hospital productivity. To do this, researchers typically choose a "target outcome" – the mortality of a patient, say – and compare it to the inputs (expenditure) associated with treating a certain condition. Figure 1 presents what is probably the modal example: one-year mortality outcomes for patients who suffer an acute myocardial infarction (AMI), popularly known as a heart attack. For each U.S. hospital that treated at least 200 AMI patients aged 65 or older between 2007 to 2009, the figure plots 1-year mortality versus expenditures.

There are a few features of this example that are noteworthy and to which I return. First, AMI is the condition most studied not only because cardiovascular disease is a leading cause of death but also because patients are nearly always treated for AMI at the same hospital where they are taken to an emergency room. This makes it easy to assign patients – and the costs their treatment incurs – to particular hospitals. Patients who suffer from other conditions might be treated at multiple hospitals. Second, short-term mortality is often the target outcome because data are available, it is very accurately measured, and reducing it is clearly a goal of AMI

[1] This chapter was written by Douglas Staiger in collaboration with Caroline Hoxby and Kevin Stange. Staiger presented slides at the NBER conference. He also engaged in a lively question and answer session. Using the slides and their notes from the session, Hoxby and Stange produced a manuscript for readers. They did not introduce ideas not already in the Staiger presentation, but they did amplify and clarify points to make them accessible to readers. Staiger then revised the manuscript, producing this chapter.

treatment. Third, both mortality and expenditures are *risk-adjusted*. This means that the researcher has controlled for characteristics of patients that were present when they arrived at the hospital. For instance, a patient might be smoker, be obese, or have concurrent diabetes. Since these conditions might make treating AMI harder and might affect mortality regardless of AMI, the researcher would not want to attribute their effects to the hospital. Otherwise, productivity would be overstated for hospitals that draw upon an unusually healthy population and *vice versa*.

The figure shows that there is large variation in risk-adjusted expenditure across hospitals. Hospitals in the highest decile spend more than $50,000 per AMI patient while those in the bottom decile spend only $35,000. There is also large variation in mortality across hospitals. Risk-adjusted one-year mortality ranges from 25% to 38% for hospitals at, respectively, the 90th and 10th percentile. Interestingly, mortality and spending are not highly correlated, implying substantial differences in *productivity* of hospitals in treating AMI patients. Some hospitals – those in the lower left quadrant – appear to be very productive. Their patients were given low cost treatment but nevertheless have low mortality rates. Other hospitals – those in the upper right quadrant – have low productivity, with high costs and high mortality. If the risk adjustments work as intended, these productivity differences are real and do not simply reflect the hospitals' different patient populations.

What have policy makers taken away from evidence such as that shown in Figure 1? They have become deeply curious about hospital and physician practices that might explain such large variation in productivity. They hope that researchers will be able to identify practices that, if adopted, would improve the low productivity hospitals. The following statement from former Director of the Office of Management and Budget and Congressional Budget Office Peter Orszag exemplifies this curiosity:

> *If we can move our nation toward the proven and successful practices adopted by lower-cost areas and hospitals, some economists believe health-care costs could be reduced by 30% -- or about $700 billion a year – without compromising the quality of care. (WSJ, May 15, 2009).*

III. Parallels between health care and higher education

Several of the key challenges to measuring productivity in health care also prevail in higher education. These include multiple outcomes, selection, the multi-product nature of health care providers, and the attribution problem.

A. Multiple outcomes

In the example, the outcome studied was one-year mortality: the fraction of patients treated for AMI who survived the first year after treatment. We might easily get health practitioners and policy makers to agree that it was reasonable to focus on this outcome rather than on, say, long-term mortality, morbidity, functional mobility, or various measures of quality of life. The ease of agreement for AMI does not imply, however, that it would *generally* be easy to obtain such widespread agreement. Rather, AMI is the modal example in part *because* agreement is easy. For other conditions, different people would put different weights on the multiple outcomes affected by treatment.

The key point is that, even for AMI, there is no correct target outcome. There is also no correct set of weights that we could use to form an index based on multiple outcomes. Choosing a target outcome or choosing index weights is inherently a value-laden decision: statistics do not help us. Rather, the choice is inevitably a reflection of our preferences and subjective judgements, not an objective truth.

With this in mind, what are some lessons from health care that apply to higher education?

First, if researchers decide to prioritize one outcome as the target or "gold standard" outcome, this choice will drive everything. A target outcome sends a message to patients, providers, and staff about institutional mission and priority. Hospital leadership will guide providers and staff to focus on the target outcome, often to the exclusion of other objectives. In extreme cases or where the target outcome is easily manipulated, there may be unintended consequences such as altering diagnoses (so that only certain patients count toward the measured outcome) or cherry-picking patients who are healthier than their risk score would suggest.

Second, the choice of a target outcome is crucial even if it is not directly used to measure productivity but instead guides how to use other indicators. For instance, in health care, indicators other than mortality are often used because they are available more quickly and are therefore more useful for immediate feedback. These include indicators of hospital use (patient volume, for example), process (use of "best practices"), and proximate outcomes (infection rates and one-month hospital readmission rates, for example). However, indicators are often selected or given weights in a composite index based on how highly correlated they are with the target outcome. As a result, the target outcome remains a driving force.

Thus, the first lesson from medicine for higher education is that the choice of a target outcome is likely to be highly consequential. Policy leaders and researchers ought to think through the decision of whether to choose a target at all. Several chapters in this volume (Hoxby; Minaya and Scott-Clayton; Riehl, Saavedra, and Urquiola; Carrell and Kurleander) demonstrate that while the multiple obvious outcomes in higher education (graduation rates, learning, public service, innovation, short-term earnings and employment, long-term earnings and employment, etc.) *are* correlated, they are not so correlated that privileging one outcome would not have the effect of undercutting other objectives.

B. Selection

Selection poses a significant challenge to estimating hospital productivity because the sorting of patients to hospitals is not random. Some hospitals, because of their specializations or unusual resources, are destination facilities for patients who are especially ill. The Mayo Clinic and top research university hospitals are examples. Other hospitals receive unusually healthy or ill patients simply because of their location. A hospital located in a poor area is likely to receive more impoverished patients, for instance.

Figure 2 provides an example. As a proxy for whether a hospital is using best practices, the figure uses the probability of Beta Blocker treatment among AMI patients. (This is a popular proxy for best practice because Beta Blockers are widely regarded as a highly effective, low-cost treatment.) What the figure shows, however, is that this best practice proxy is highly correlated with patients' income. The correlation is so high, 0.59, that it could not possibly be generated by random sorting of patients to hospitals. In an environment with such obvious selection, measuring hospitals' productivity is hard because we need to separate the contribution of hospitals to outcomes from the contribution of patients' own characteristics to their outcomes. Recall our motivating AMI example (Figure 1). Do the hospitals in the lower left quadrant appear to the especially productive because they use effective, inexpensive Beta Blockers? Or, are they in the lower left quadrant because their patients have higher incomes? The raw data cannot answer these questions.

Fortunately, in health care it appears that by applying risk adjustment procedures to raw data, we can remedy much of the potential selection bias. In theory, any condition with which a patient arrives at the hospital door should be categorized for use in risk adjustment. In fact, the coding of hundreds of risks by the Centers for Medicare and Medicaid Services (CMS) is well regarded and widely used. An example of a risk is "F10.20 Alcohol Dependence, uncomplicated." The only obvious risk factors that are not widely used to address selection are the socio-demographics of a hospital's patient population. CMS discourages their use despite evidence that it is harder to treat poor and minority patients.

How do we know that risk adjustment largely remedies selection in health care? We compare risk-adjusted estimates with those derived from experiments where patients are randomly or quasi-randomly assigned to hospitals so that selection is not an issue.[2] We compare risk-adjusted estimates with direct observation of surgical technical quality.[3] We compare risk-adjusted estimates with estimates where a researcher is able, owing to especially rich and complete data, to control for *comprehensive* clinical information at the patient level.[4] In all of these cases, standard risk-adjusted estimates of hospital performance compare favorably to credibly causal estimates.

[2] See Doyle et al., 2010; Doyle, 2011; Doyle et al., 2014.

[3] See Birkmeyer et al., 2013.

[4] See Dimick et al., 2008; Dimick et al., 2010; McClellan & Staiger, 2000; Morales et al., 2005; Skinner et al., 2005.

Can we extrapolate to higher education the finding that risk adjustment largely remedies selection in health care? Not obviously. On the one hand, in both hospitals and colleges, much of selection arises because of geography (people use nearby institutions) or self-selection (better informed and higher income people may seek out better institutions). Another similarity is that selection issues are particularly problematic for the most resource-intensive institutions in both health care (the Mayo Clinic, for instance) and higher education (Harvard University, for instance). A difference, though, is in the direction of selection for the most resource-intensive institutions: The most-resourced hospitals see the least healthy patients, while the best-resourced universities serve the most able students. Consequently, if we fail to account for non-random selection, the most-resourced hospitals will appear to be low-performing while the most-resourced universities will appear to be high-performing. Another difference between health care and higher education is that hospitals do not explicitly practice selective admission while many postsecondary institutions do. However, explicitness seems likely to make selection *easier* to remedy in higher education. A researcher may know for certain or at least have a very good idea of the factors that a college is weighing in the admissions process and can use this information to account for the selection generated by the admissions process.

Summing up, the second lesson from medicine for higher education is that researchers ought not to assume that selection is so unremediable that it is pointless to work on developing the best possible adjustment procedures for pre-college factors like high school achievement and family background. In health care, research devoted to risk adjustment has borne fruit. To validate their adjustment procedures, higher education researchers should compare their estimates to estimates generated by policy experiments or quasi-experiments like discontinuities in admission criteria. While experiments and quasi-experiments are not common in higher education, they are sufficiently common for such validation exercises, which have proven so useful in health research.

C. Multiproduct issues

Hospitals provide multiple service lines, delivered by different departments: oncology, cardiology, infectious disease, and so on. Each department also employs an array of procedures: surgery, biopsy, blood testing, radiology, etc. In the language of economics, hospitals are multi-product organizations. Hospitals also serve multiple populations, most notably patient populations whose risk profiles differ. If each hospital were equally productive in all its departments, procedures, and patient populations, then it would not matter which we examined when evaluating a hospital.

But, in fact, research suggests that hospitals are not equally productive in all their service lines. Consider patients treated in two important service lines: AMI and hip fractures. Figure 3 plots the risk-adjusted one-year mortality rate for AMI patients against the one-mortality rate for hip fracture patients. Both measures are for patients aged 65 or older who were treated at hospitals that saw at least 200 such patients. Hospital performance across these two service lines is only modestly correlated with a correlation coefficient of 0.30. Other research indicates that most of the variation in productivity within hospitals across departments comes from variation in patient outcomes (the numerator) rather than patient costs (the denominator). Still further research suggests that variation in productivity is mainly *across* departments, not mainly (i) within departments across procedures or (ii) within departments across patients' risk profiles. In other words, departments seem to have integrity as service providers and give high or low quality service regardless of their hospital's cost structure, procedural units, and patient risk profiles.

For hospitals, this all suggests a need to measure productivity separately by service line (department). It also indicates that evaluating a hospital based on a single service line (oncology, say) is likely to be problematic. Narrow evaluation might encourage a hospital to reallocate fungible resources (nurses, laboratory time, etc.) from departments that are not evaluated to departments that are.

Institutions of higher education are also multi-product organizations that serve multiple populations. The typical institution supports many different departments and programs and distributes many different types of degrees. Some of the students may be undergraduates of traditional age. Others may be graduate students, professional students, or non-traditional undergraduates. Postsecondary schools also have the equivalent of procedural units that serve many departments ─ libraries, for instance.

The third lesson that higher education can learn from health care is therefore that there is no reason to assume that the evidence will show that a postsecondary institution is equally productive across all its service lines (departments, programs), student populations, and library-like procedural units. This may seem like a discouraging lesson because it implies that researchers have a formidable task ahead of them: estimating productivity for each activity at each postsecondary institution. However, it is worthwhile pointing out that health care researchers have made great progress by first focusing their attention on service lines that are important and central. By important, I mean that the service line is crucial to a hospital's identity. By central, I mean that the service line deals with broad swath of the patient population and uses many procedural units. If a service line is central, its productivity is less easily manipulated by the hospital moving resources around (the multi-tasking problem).

For instance, the fact that the most examined department is cardiology is not an accident. As emphasized above, cardiovascular disease is a leading cause of death so this area is important to most hospitals. Moreover, cardiovascular disease is not rarified or confined to some minority of the potential patient population. The cardiology department also draws upon many procedural units. Cardiology is therefore central.

By parallel logic, higher education researchers might first focus on undergraduate education because it tends to be important and central. That is, the quality of its undergraduate program is key to most (though not all) institutions' identities, and undergraduates draw upon a wide of departments and procedural units (libraries, etc.). Researchers might secondarily focus on the high profile professional and doctoral programs that define research universities.

D. Attribution

Patients often interact with multiple hospitals as well as other health care providers when they are being treated for a condition. Although we know which procedures and which costs are attributable to each provider, patients' outcomes (their one-year mortality, for example) cannot so easily be assigned to providers. Their outcomes are presumably due to the entire sequence of care. Moreover, it is not obvious that a provider's responsibility is proportional to its share of costs. Changing the quality of even a single procedure could be consequential if other procedures are endogenous to it. For instance, if cardiac catheterization were poorly performed, all of a patient's subsequent treatment for heart disease might be less effective. Thus, when attempting to measure productivity in health care, we often face the question of how to attribute patients' outcomes to individual hospitals or other providers.

Health care researchers have found two ways to deal with this problem. First, they often focus on conditions, such as AMI, where the attribution problem is minimal for technical reasons. That is, when people suffer heart attacks, they are usually taken to the closest hospital with cardiac capacity, and they are treated there until released. Second, health care researchers often define health "episodes" that begin with a diagnosis or event (such as a stroke) and then attribute all or most care within the episode to the hospital in which treatment began. The logic is that the initial hospital made choices to which all subsequent treatment (in the episode) is endogenous. A person may have multiple health episodes in his life.

The first of these solutions, focusing on situations where the attribution problem is minimal, does not seem helpful for higher education, where the attribution problem occurs because students (i) take classes at various institutions while pursuing the same degree and (ii) engage in degree programs serially, with each degree at a different institution. One-third of students transfer institutions at least once within their first six years of college and before receiving a bachelor's degree and nearly one-sixth transfer more than once.[5] Or consider people who earn an associate's degree at a community college, a baccalaureate degree at a (different) four-year college, a master's degree at a third institution, and a professional degree at yet a fourth institution. To which institution should their post-professional-degree outcomes, such as earnings, be attributed? If researchers were to exclude all students whose education spanned multiple institutions, the exclusion would be highly non-random and introduce bias. There is no parallel to AMI.

[5] National Student Clearinghouse, 2015. "Signature Report: Transfer & Mobility: A National view of Student Movement in Postsecondary Institutions, Fall 2008 Cohort" downloaded from https://nscresearchcenter.org/wp-content/uploads/SignatureReport9.pdf

The second solution is more promising: attribute productivity to the initial institution in an educational episode where an episode is defined by fairly (though not entirely) continuous enrollment. People might still have multiple episodes if they, for instance, attained a baccalaureate degree between age 18 and age 24 and then, after an interval of more than a decade, enrolled in a master's degree program.[6] This approach is exemplified by Hoxby's chapter in this volume.

A fourth lesson that higher education can take from health care is therefore that attribution issues, while important, can be overcome by treating educational episodes that span multiple institutions as the object of interest. It may not be desirable – or feasible – to try to separate the individual contribution of a community college from the four-year institution it feeds. It should be noted that identifying health episodes spanning multiple providers requires patient-centric data that track patients across these providers. Similarly, identifying education episodes spanning multiple institutions requires student-central data.

IV. Lessons for measuring and using productivity in higher education

The experience of measuring productivity in health care offers four main lessons to similar efforts in higher education. First, the choice of a target outcome is likely to be highly consequential. Policy leaders and researchers ought to think through the decision of whether to choose a target at all or how multiple targets should be combined. Graduation, alumni earnings and employment, innovation, student learning – these are all plausible objectives of postsecondary institutions and systems, but giving priority to one may generate neglect of the others. Second, although selection issues are important, adjusting for selection may be successful if rich enough controls are available. In addition, selection-adjustment should be validated by experimental or quasi-experimental evidence. Third, institutions are unlikely to be equally productive across all their service lines and populations. Initial productivity measurement should focus on service lines that are important and central, such as undergraduate education. Finally, attribution issues can be overcome by treating educational episodes that span multiple institutions as the object of interest, attributing outcomes to the initial institution.

I conclude with two broad lessons about the use of productivity measures in health care that may also inform how they are used in higher education.

[6] A third possibility is suggested by value-added research in elementary and secondary education. A few researchers have attempted to identify the long-term value-added (to adult earnings, say) of each teacher in a succession of teachers who instruct a student. As an econometric matter, such identification is possible so long as students' teacher successions sufficiently overlap. The Carrell and Kurlaender study in this volume illustrates this approach which works, in their case, because many California students attend overlapping community colleges and California State Universities. However, this solution is often infeasible in higher education because students are not channeled so neatly through a series of institutions as through a series of primary and secondary teachers: the teachers available in a school in a grade are much more limited than the institutions among which students can choose. Postsecondary students are also not channeled so neatly through a series of grades: they can exit, get labor market experience between periods of enrollment, choose multiple degree paths, and so on.

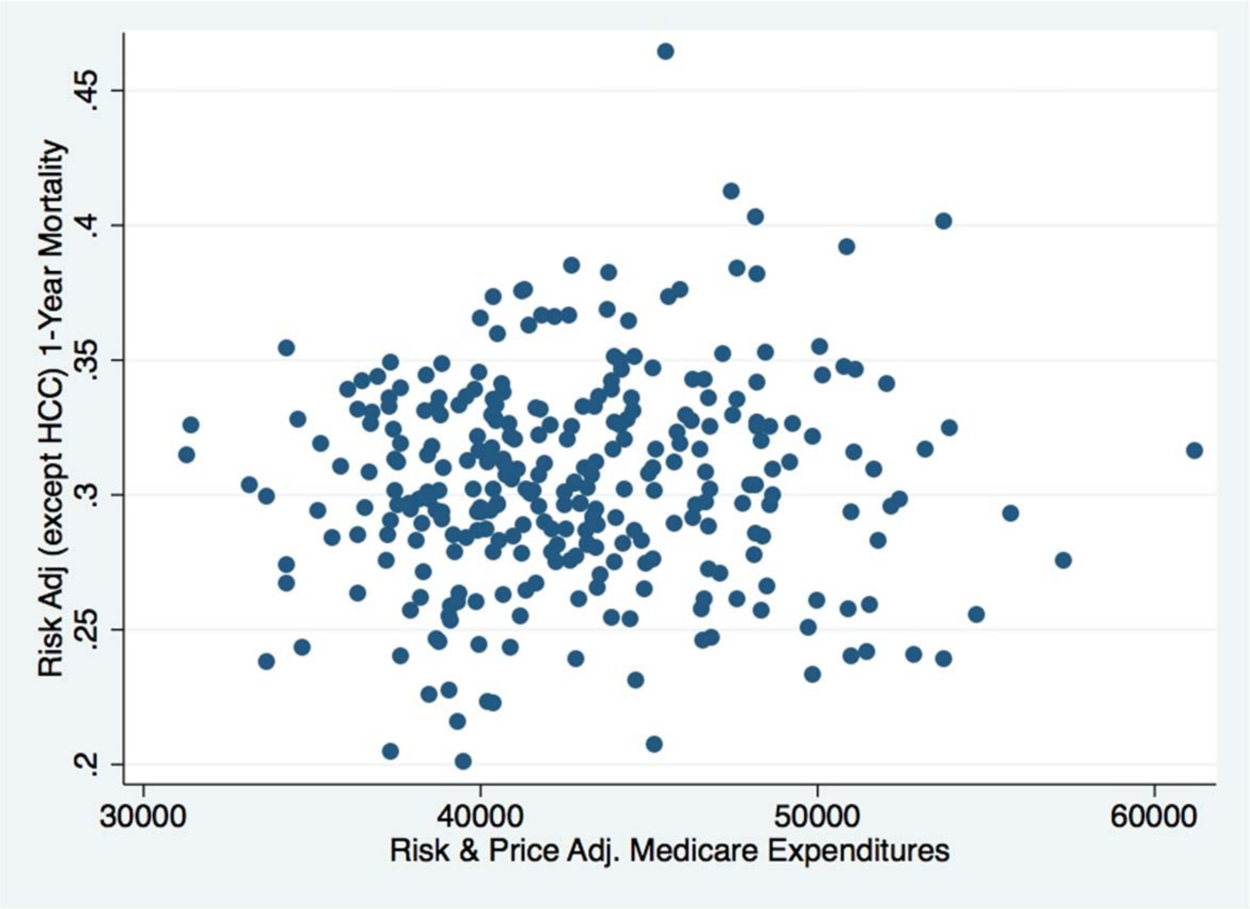
First, productivity measures have multiple uses and there ought to be a match between the productivity measure and the use made of it. Productivity measures in health care have been used to inform patients who are trying to choose a provider, make providers accountable for health outcomes and costs, and to provide timely feedback so that providers continuously improve. These different uses require different measures. For instance, patients may care about how a provider will affect their health and the costs they themselves will pay, but they may be uninterested in costs paid by insurers. Patients may also care more about, say, the treatment experience as a whole while policy makers care more about mortality or disability. These differences may explain why patients seem to make little use of hospital "report cards," while insurers make considerable use of them to direct patients toward providers that appear to be more productive. If the productivity measures published in the report cards are only those requested by insurers and policy makers, it should no surprise that patients ignore them. As another example, hospitals that are trying to adjust their processes to improve treatments require productivity measures that are very timely. They may be willing to sacrifice accuracy and knowledge of long-term benefits so that they can adjust processes in real time. Patients and policy-makers presumably weigh accuracy more and timeliness less.

Second, stakeholder buy-in is important if we are to see university leaders (especially) but also students and policy makers take productivity measures to heart. They will not use them to improve decision-making if they find them unconvincing. Buy-in is especially important in hospitals and postsecondary institutional because they are inherently decentralized organizations where much expertise resides in departments or even in individual physicians or faculty. Crucial testing, treatment/curricular, and staffing decisions must inevitably be delegated to those with the expertise. Thus, productivity measures will only be used well if they truly respected by individuals and units throughout the health care/higher education organization. For instance, suppose that university leaders think that initial earnings are beyond their control but agree that learning (as measured by an exit exam, say) is within their control. Suppose furthermore that learning is more correlated with long-term earnings and employment outcomes, which university leaders care about, then are initial earnings. In such a case, productivity measures must include learning-based outcomes if they are to enjoy actual use by leaders. In health care, efforts to measure productivity and have the measures actually inform stakeholders' decisions were only successful when researchers sought input from those same stakeholders. This is a lesson that surely applies to higher education.

NOTE TO DOUG: We considered another subsection on "Outcome timing," but instead wrapped this into the discussion of "Multiple outcomes" and "Uses of productivity measures." This was in part because we did not have any great examples of how health care has thought about or confronted the outcome timing issue (e.g. looking at short-term vs. long-term mortality). This topic could be a new subsection III.E. if we thought there was enough additional content to pull it out separately.

Figure 1. An Illustrative Example From Healthcare

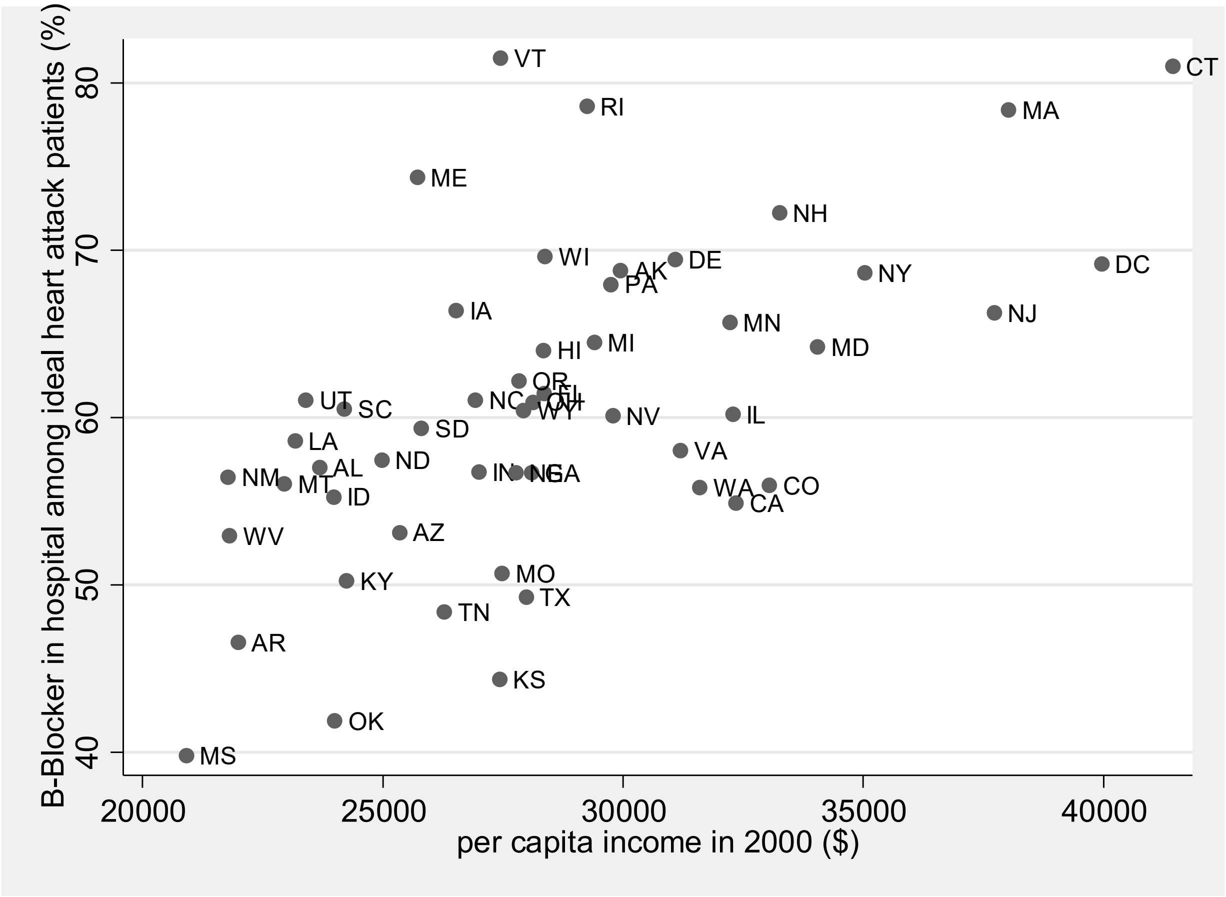
Risk+Price Adjusted 1-Year Expenditures and Mortality by Hospital



Note: Sample limited to hospitals with at least 200 AMI patients age 65+; 2007-09. Source: Author's analysis of Medicare claims data.

Figure 2. Per Capita Income and Beta Blocker Use in the Hospital Among Ideal Heart Attack Patients

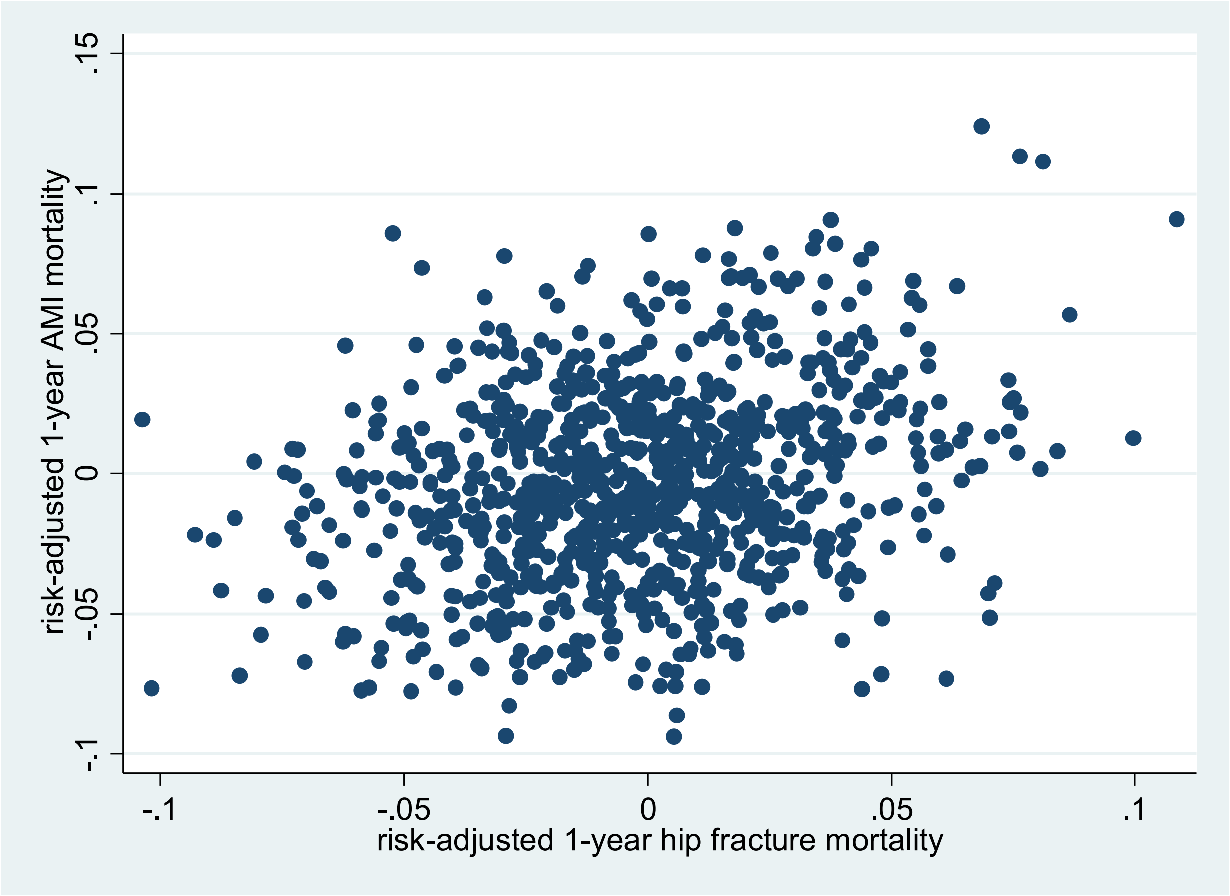
(correlation=.59)



Note: Data on Beta Blocker use from Cooperative Cardiovascular Project, 1994-1995

Figure 3. Correlation Across Departments

Correlation (0.3) in Hospital Mortality Rates For AMI and Hip Fracture Patients



Note: Sample limited to hospitals with at least 200 AMI & hip patients age 65+; 2000-02.