PHYSICIAN BELIEFS AND PATIENT PREFERENCES: A NEW LOOK AT REGIONAL VARIATION IN HEALTH CARE SPENDING

David Cutler
Harvard University and National Bureau of Economic Research

Jonathan Skinner
Dartmouth College, Dartmouth Institute for Health Policy & Clinical Practice, and National Bureau of Economic Research

Ariel Dora Stern
Harvard Business School

David Wennberg
Dartmouth Institute for Health Policy & Clinical Practice

Abstract

There is considerable controversy about the causes of regional variations in health care expenditures. We use vignettes from patient and physician surveys linked to Medicare expenditures at the Hospital Referral Region to test whether patient demand-side factors or physician supply-side factors explain regional variations in spending. We find patient demand is relatively unimportant in explaining variation in spending after accounting for physician beliefs. Physician organizational factors matter, but the single most important factor is physician beliefs about treatment: 35 percent of end-of-life spending, and 12 percent of U.S. health care spending, are associated with physician beliefs unsupported by clinical evidence.

Corresponding Author: Ariel Dora Stern
Harvard Business School
Morgan Hall 442
Boston, MA 02163
Phone: (617) 495 2332
astern@hbs.edu

Comments by Amitabh Chandra, Elliott Fisher, Mike Geruso, Tom McGuire, Joshua Gottlieb, Nancy Morden, Allison Rosen, Gregg Roth, Pascal St.-Amour, Victor Fuchs, and seminar participants were exceedingly helpful. We are grateful to F. Jack Fowler and Patricia Gallagher of the University of Massachusetts Boston for developing the patient and physician questionnaires. Funding from the National Institute on Aging (T32-AG000186-23 and P01-AG031098 to the National Bureau of Economic Research and P01-AG019783 to Dartmouth) and LEAP at Harvard University (to Skinner) is gratefully acknowledged. De-identified survey data are available at http://tdicauses.dartmouth.edu.

I. Introduction

Regional variations in rates of medical treatments are large in the United States and other countries (Skinner, et al., 2012). For example, in the U.S. Medicare population over age 65, 2011 price-adjusted per-patient Medicare expenditures by hospital referral region (HRR) ranged from $6,876 in La Crosse, Wisconsin, to $13,414 in Miami, with most of the variation unexplained by regional differences in patient illness or poverty (Institute of Medicine, 2013).

What drives such variation in treatment and spending? One possibility is patient demand. Most studies of variations have been conducted in environments where all patients have a similar and generous insurance policy, so price differences and income effects are likely to be small. Still, heterogeneity in patient preferences for care may play a role. In very acute situations, some patients may prefer to try all possible measures, while others may prefer palliation and an out-of-hospital death. If people who value and demand life-prolonging treatments live in the same areas, patient preference heterogeneity could lead to regional variation in equilibrium outcomes (Anthony, et al., 2009; Baker, et al., 2014; Mandelblatt, et al., 2012).

Another possible source of variation arises from the supply side. In "supplier-induced demand," a health care provider shifts a patient's demand curve beyond the level of care that the (fully informed) patient would otherwise want. This would be true in a principle-agent framework (McGuire and Pauly, 1991) if prices are high enough or income is scarce. While physician utilization has been shown to be sensitive to prices (Jacobson, et al., 2006; Clemens and Gottlieb, 2014), it would be difficult to explain the magnitude of Medicare variations using prices alone, since federal reimbursement rates do not vary greatly across areas.

Variations in supply could also occur if physicians respond to organizational pressure or peer pressure to perform more procedures, or maintain differing beliefs about appropriate treatments – particularly for conditions where there are few professional guidelines (Wennberg, et al., 1982). If this variation is spatially correlated – for example, if physicians with more intensive treatment preferences are more likely to hire other physicians with similar practice styles – the resulting regional differences in beliefs could explain regional variations in equilibrium spending.

It has proven difficult to separately estimate the impact of physician beliefs, patient preferences, and other factors as they affect equilibrium healthcare outcomes, largely because of challenges in identifying factors that affect only supply or demand (Dranove and Wehner, 1994). Finkelstein, et al. (2014) get around many of these challenges by following Medicare enrollees who moved to more or less expensive regions, allocating roughly half of the difference in spending to supply, and half to demand. However, their measure of demand includes unmeasured health status, and thus cannot identify the influence of patient preferences *per se*. Nor can they easily discern the causes of these supply-side effects; are they income effects, substitution effects, or some combination of the two and possibly other factors?

In this paper, we use "strategic" surveys of physicians and patients (in the sense of Ameriks et al., 2011) to sidestep issues of risk-adjustment and endogeneity in estimating a unified model of overall regional Medicare expenditures. Physician incentives and beliefs are captured using detailed, scenario-based surveys that present physician-respondents with questions about their financial and practice organization, and with vignettes about how the physician would manage elderly individuals with specific chronic health conditions and a given medical and treatment history. Vignettes have been shown to predict actual physician behavior in various clinical settings (e.g. Peabody, et. al, 2000 and 2004; Mandelblatt, et al., 2012; Evans, et al., 2015). In prior work, the physician surveys have been used by Sirovich, et al. (2008) and

Lucas et al. (2010) to find that Medicare spending predicted physician treatment intensity, but these studies did not adjust for patient demand, nor could they answer the question of whether (and which) supply side factors explain observed variation in Medicare expenditures.[1]

Patient preferences are measured by a survey of Medicare enrollees age 65 and older asking about whether they would want a variety of aggressive and/or palliative care interventions. Survey questions have also been demonstrated to predict individual care-seeking behavior (Mandelblatt, et al., 2012; Anthony, 2009). Previous studies using this survey (Anthony, et al., 2009; Barnato, et al., 2007; Baker, et al., 2014) found a small but discernible influence of patient preferences on regional variations in health care spending. Baker et al. (2014) attempted to control for supply-side effects by using measures such per capita levels of hospital beds or physicians. However, these equilibrium quantities are endogenously determined by both supply and demand, leading to inconsistent estimates.

We first develop an equilibrium model that specifies health care intensity (as measured by expenditures) as a function of a variety of factors specific to health care providers' and patients' preferences. To estimate this requires a multivariate model in which supply and demand factors are on the right-hand side of the equation, and overall intensity (or expenditures) are on the left. The downside to estimating the theoretical model is that in order to ensure sufficient sample size within each HRR, we are restricted to using the 74 largest Hospital

[1] In these studies, because spending was on the right-hand side of the equation, and physician survey responses on the left, one cannot estimate whether survey responses explain regional variations in spending, nor can the different factors (e.g., financial incentives, vignette responses) be considered in a multivariate framework.

Referral Regions (HRRs) in the United States, accounting for about half of the elderly population.[2]

Our model characterizes physicians along two non-exclusive dimensions: those who consistently and unambiguously recommended intensive care beyond those indicated by current clinical guidelines (which we refer to as "cowboys"), and those who consistently recommended palliative care for the very severely ill (which we refer to as "comforters"). We can explain over half of the variation in end-of-life spending across areas by knowing only these two measures of physician treatment intensity, as well as the frequency with which physicians recommend that their patients return for routine office visits. Our results are consistent with those in Finkelstein, et al. (2014) whose estimated supply-side variations are similar to (or a bit larger than) those reported below.[3] Conditional on the supply side, we find that demand-side Medicare patient preferences explains about 8 percent or less of total spending, a result also consistent with previous studies (Anthony, et al., 2009; Barnato, et al., 2007; Baker, et al., 2014; and Gogineni, et al., 2015).

Previous studies have not been able to explain why some physicians are cowboys, some are comforters, and others are neither. We use the full sample of surveyed physicians to examine factors associated with the likelihood of being a cowboy or a comforter. We find that only a small fraction of physicians claim to have made recent decisions as a result of purely financial considerations. We also find that "pressure to accommodate" either patients (by providing treatments that are not needed) or referring physicians' expectations (doing procedures to keep them happy and meet their expectations) have a modest but significant relationship with

[2] Ideally, we would have linked the physician surveys directly to how they treated patients. We were precluded from doing so, however, because we did not have the permission of the physicians to link their survey responses to claims data.

[3] Grytten and Sørensen (2003) also find a large role for supply-side factors.

physician beliefs about appropriate care. While many physicians do report making interventions as a result of malpractice concerns, these responses do not explain the residual variation in treatment recommendations.

Ultimately, the largest degree of residual variation appears to be explained by differences in physician beliefs about the efficacy of particular therapies. Physicians in our data have starkly different views about how to treat the same patients. These views are not strongly correlated with demographics, financial incentives, background, or practice characteristics, and are often inconsistent with evidence-based professional guidelines for appropriate care. As much as 35 percent of end-of-life Medicare expenditures, and 12 percent of overall Medicare expenditures are explained by physician beliefs that cannot be justified either by patient preferences or by evidence of clinical effectiveness.

II. A Model of Variation in Utilization

We develop a simple model of patient demand and physician supply. The demand side of the model is a standard one; the patient's indirect utility function is a function of out-of-pocket prices (p), income (Y), health (h), and preferences for care (η); $V = V(p, Y, h, \eta)$. Solving this for optimal intensity of care, x, yields x^D. As in McGuire (2011), we assume that x^D is the fully informed patient's demand for the quantity of care or procedures prior to any demand "inducement."

On the supply side, we assume that physicians seek to maximize the perceived health of their patient, s(x), by appropriate choice of inputs x, subject to patient demand (x^D), financial considerations, and organizational factors. Note that the function s(x) captures both patient survival and patient quality of life, for example as measured by quality-adjusted life years (QALYs).

Individual physicians are assumed to be price-takers (after their networks have negotiated prices with insurance companies), but face a wide range of reimbursement rates from private insurance providers, Medicare, and Medicaid. The model is therefore simpler than models in which hospital groups and physicians jointly determine quantity, quality, and prices, (Pauly, 1980) or in which physicians exercise market power over patients to provide them with "too much" health care (McGuire, 2011). Following Chandra and Skinner (2012), we write the physician's overall utility as:

(1) $$U = \Psi s(x) + \Omega(W + \pi x - R) - \phi(|x - x^D|) - \varphi(|x - x^O|)$$

where Ψ is perceived social value of improving health, Ω is the physician's utility function of own income, comprising her fixed payment W (a salary, for example) net of fixed costs R, and including the incremental "profits" from each additional test or procedure performed, π.[4] The sign of π depends on the type of procedure and the payment system a physician faces.

The third term represents the loss in provider utility arising from the deviation between the quantity of services the provider actually recommends (x) and what the informed patient demands (x^D). This function could reflect classic supplier-induced demand – from the physician's point of view, x^D is too low relative to the physician's optimal x – or it may reflect the extent to which physicians are acting as the agent of the (possibly misinformed) patient, for example when the patient wants a procedure that the physician does not believe is medically appropriate. The fourth term reflects a parallel influence on physician decision-making exerted by organizational factors that do not directly affect financial rewards, such as (physician) peer pressure.

[4] We ignore capacity constraints, such as the supply of hospital or ICU beds.

The first-order condition for (1) is:

(2) $$\Psi s'(x) = -\Omega'\pi + \phi' + \varphi' \equiv \lambda$$

Physicians provide care up to the point where the choice of x reflects a balance between the perceived marginal value of health, $\Psi s'(x)$, and factors summarized by λ: (a) the incremental change in net income π, weighted by the importance of financial resources Ω', (b) the incremental disutility from moving patient demand away from where it was originally, ϕ', and (c) the incremental disutility from how much the physician's own choice of x deviates from her organization's perceived optimal level of intervention, φ'.

In this model,[5] there are two ways to define "supplier-induced demand." The broadest definition is simply the presence of any equilibrium quantity of care beyond the level of the *ex ante* preferences of an informed patient, i.e. $x > x^D$. This is still relatively benign, as the marginal value of this care may still be positive. Supplier-induced demand could more narrowly be defined as $s(x) - s(x^D) \leq 0$; for additional care provided at the margin, patients gain no improvement in health outcomes and may even experience a decline in health or a significant financial loss. Importantly, both of these definitions are ambiguous about the question of physician *knowledge* of inducement beyond clinically appropriate levels. That is, a physician with strong (but incorrect) beliefs may over-treat her patients, even in the absence of financial or organizational incentives to do so.

To develop an empirical model, we adopt a simple closed-form solution of the utility function for physician i:[6]

(1′) $$U_i = \Psi s_i(x_i) + \omega[W_i + \pi_i x_i - R_i] - \frac{\phi}{2}(x_i - x_i{}^D)^2 - \frac{\varphi}{2}(x_i - x_i{}^O)^2$$

---

[5] A more general model would account for the patient's ability to leave the physician and seek care from a different physician, as in McGuire (2011).

[6] We are grateful to Pascal St.-Amour for suggesting this approach.

Note that ω/Ψ reflects the relative tradeoff between the physician's income and the value of improving patient lives, and thus might be viewed as a measure of "professionalism," as in Campbell, et al. (2010). The first-order condition is therefore:

$$(2') \qquad \Psi s_i'(x_i) = \lambda \equiv -\omega\pi_i + \phi(x_i - x_i{}^D) + \varphi(x - x_i{}^O)$$

Figure 1 shows $\Psi s'(x)$ and λ. Note that λ is linear in x with an intercept equal to $-(\omega\pi_i + \phi x_i{}^D + \varphi x_i{}^O)$. Note also the key assumption that patients are sorted in order from most appropriate to least appropriate for treatment, thus describing a downward sloping $\Psi s'(x)$ (marginal utility of treatment) curve. The equilibrium is where $\Psi s'(x) = \lambda$, at point A. A shift in the intercept, which depends on reimbursement rates for procedures π, taste for income ω, regional demand x^D, and organizational or peer effects x^O, would yield a different λ^*, and hence a different utilization rate. However, all of these factors affect the intensity of treatments via a movement *along* the marginal benefit curve, $\Psi s'(x)$.

Alternatively, it may be that $s_i'(x)$ differs across physicians – i.e. physician productivity differs, rather than physician constraints. For example, if $g_i'(x) = \alpha_i + s'(x)$, where $s'(x)$ is average physician productivity and α varies across regions, this would be represented as a shift in the marginal benefit curve. Point C in Figure 1 corresponds to greater intensity of care than point A and arises naturally when the physician is *or just believes* she is more productive. For example, heart attack patients experience better outcomes from cardiac interventions in regions with higher rates of revascularization, consistent with a Roy model of occupational sorting (Chandra and Staiger, 2007). Because patients in regions with high intervention rates benefit differentially from these interventions, this scenario does not correspond to the narrow definition of "supplier-induced demand."

The productivity shifter α_i may also vary because of "professional uncertainty" – a situation in which the physician's perceived α_i differs from the true α_i (Wennberg, et al., 1982). Physicians may be overly optimistic with respect to their ability to perform procedures, leading to expected benefits that exceed actual realized benefits. Baumann, et al. (1991) have documented the phenomenon of "macro uncertainty, micro certainty" in which physicians and nurses exhibit overconfidence in the value of their treatment for a specific patient (micro certainty) even in the absence of a general consensus as to which procedure is more clinically effective (macro uncertainty). Ransohoff, et al. (2002) has noted a further psychological bias towards more aggressive treatment: If the patient gets better, the physician gets the credit, but if the patient gets worse, the physician is able to say that she did everything possible.

To see this in Figure 1, suppose the actual benefit is $s'(x)$ but the physician's perceived benefit is $g'(x)$. The equilibrium is point D: the marginal treatment harms the patient, even though the physician believes the opposite, incorrectly believing they are at point C. In equilibrium, this supplier behavior would appear consistent with classic supplier-induced demand, but the cause is quite different.

Empirical Specification. To examine these theories empirically, we consider variation in practice at the regional level (for reasons explained below) but adjusting for health status, h. Taking a first-order Taylor-series approximation of equation (2′) for region i yields a linear equation that groups equilibrium outcomes into two components, demand factors Z^D and supply factors Z^S:

(4) $$x_i = \bar{x} + Z_i^D + Z_i^S + \varepsilon_i.$$

The demand-side component is:

(5) $$Z_i^D = \frac{\phi}{\mathrm{M}}(x_i^D - \bar{x}^D)$$

where $M = -\Psi s''(\bar{x}) + \phi + \varphi$. This first element of equation (5) reflects the higher average demand for health care, multiplied by the extent to which physicians accommodate that demand, ϕ. The supply side component is:

$$(6) \qquad Z_i^S = \frac{1}{M}\{\omega\Delta\pi_i + \pi\Delta\omega_i + \varphi({x_i}^O - \bar{x}^O) + \Psi\Delta\alpha_i\}$$

The first term in equation (6) reflects how the difference in profits in region i vs. the national average ($\Delta\pi$) affects utilization. The second term reflects the extent to which physicians weigh income more heavily. The third term captures organizational goals in region i relative to national averages $({x_i}^O - \bar{x}^O)$. The final term captures the impact of different physician beliefs about productivity of the treatment ($\Delta\alpha_i$); this term shifts the marginal productivity curve.

Equation (4) can be expanded to capture varying parameter values as well. For example, in some regions, physicians may be more responsive to patient demand (a larger ϕ_i). Such interaction effects, considered below, would reflect the interaction of supply and demand and would magnify the responses here.

III. Data and Estimation Strategy

In general, it is difficult to distinguish among various demand and supply explanations for treatment variation; even detailed clinical data reveal only a subset of what the physician knows about her patient's health and reveal virtually nothing about non-clinical drivers of patient demand for health care services. Further, patient preferences and physician beliefs about the desirability or appropriateness of different procedures are unknown in *ex post* clinical data. Ameriks, et al. (2011) was confronted with similar challenges in identifying motives for private savings, and turned to "strategic surveys," where they asked respondents specific questions about why they saved.[7]

We use physician survey vignettes to proxy for $\Psi\Delta\alpha_i/M$, and the patient surveys and vignettes to capture $\phi(x_i^D - \bar{x}^D)/M$. The use of survey vignettes is relatively novel in empirical health economics, but the validity of such studies has been demonstrated in a number of clinical settings, with vignette data on medical records found to be more accurate than data extracted from *ex post* abstraction (Peabody, et. al, 2000 and 2004).

We assume that the physician's responses to the vignettes are "all in" measures (Z_S, as in equation 6), reflecting physician beliefs as well as the variety of financial, organizational, and capacity-related constraints physicians face. In other words, we assume that if a physician were motivated entirely by financial incentives, her response to the vignettes would reflect that motivation. Alternatively, one could interpret the physician's responses to the vignettes as a pure reflection of beliefs (for example, how one might answer for qualifying boards), and not as representative of the day-to-day realities of their practice. In the appendix, we tested this alternative explanation by including the organizational and financial variables in the estimation equations in addition to the vignette estimates, but they added little explanatory power to the regression, supporting our "all-in" assumption.

In an ideal world, patient surveys would be matched with surveys from their respective physicians. Because our data do not match physicians with their own patients (nor are we permitted to match physician responses to their claims data), we instead match supply and

[7] Our approach, and Ameriks, et al. (2011), are distinct from contingent valuation studies that seek to measure the value of non-market goods such as environmental quality. Instead, we ask what individuals (physicians or patients) would do in certain well-defined scenarios.

demand at the area level using Hospital Referral Regions (HRRs).[8] In equation (4), we define x to be a regional average spending measure. Our primary measure of x is the natural logarithm of risk- and price-adjusted Medicare expenditures in the last two years of life, but we also consider several alternative measures of utilization, discussed below.

Our first estimation, based on Equation 4, asks whether area-level supply or demand factors can better explain actual regional expenditures. Our second set of estimates, based on Equation 6, then seeks to understand why physicians hold the beliefs they do. For the latter, we relate individual physician vignette responses to physicians' financial and organizational incentives. We interpret the component of vignette responses that cannot be explained by demographic, organizational or financial incentives as reflecting residual physician beliefs.

Physician Surveys. A total of 999 cardiologists were randomly selected to receive the survey. Of these, 614 cardiologists responded, for a response rate of 61 percent. Some physicians surveyed did not self-identify as cardiologists, and others were missing crucial information such as practice type, or practiced in HRRs with too few respondents to include in the analysis, leaving us a final sample of 598 cardiologists. These cardiologists practice in 74 HRRs, each of which have three or more cardiologists (as well as at least 3 patients and at least one additional PCP) represented in the survey.

The primary care physician (PCP) responses come from a parallel survey of PCPs (family practice, internal medicine, or internal medicine/family practice). A total of 1,333 primary care physicians were randomly selected to receive the survey and 973 individuals responded for a response rate of 73 percent. A total of 935 PCPs had both complete responses to the survey and practiced in HRRs with enough local patient and physician respondents to include in the final

[8] These HRRs are defined in the *Dartmouth Atlas of Health Care*, which divides the United States into 306 HRRs. Spending measures are based on area of patient residence, not where treatment is actually received.

analysis. Physician survey responses may vary systematically by demographic covariates such as age and gender. For all exercises that require aggregation of multiple physician surveys, we create demographically-adjusted HRR-level measures of physician beliefs by adjusting all responses for observed physician characteristics (race, age and sex).

Patient Survey. The survey sampling frame was all Medicare beneficiaries in the 20 percent denominator file who were age 65 or older on July 1, 2003 (Barnato et al., 2007). A random sample of 4,000 individuals was drawn; the response rate was 65 percent. Since survey responses may vary systematically by demographic covariates such as race and ethnicity; we create demographically-adjusted HRR-level measures of patient preferences by adjusting all responses for observed patient characteristics (race, age and sex).

Medicare Utilization Data. We match the survey responses with expenditure data by HRR. Our primary measure is Medicare expenditures in the last two years of life for enrollees over age 65 with a number of fatal illnesses.[9] All HRR-level measures are adjusted for age, sex, race, differences in Medicare reimbursement rates and the type of disease (including an indicator for multiple diseases). This measure implicitly adjusts for differences across regions in health status; an individual with renal failure who subsequently dies is likely to be in similar (poor) health regardless of whether she lives. End-of-life measures are commonly used to instrument for health care intensity, (e.g., Fisher et al., 2003a and 2003b; Doyle, 2011), and do not appear sensitive to the inclusion of additional individual-level risk-adjusters (Kelley et al., 2011).

A second measure of utilization captures Medicare expenditures for one year after an admission to hospital for a heart attack, or acute myocardial infarction (AMI). We further adjust

[9] These include congestive heart failure, cancer/leukemia, chronic pulmonary disease, coronary artery disease, peripheral vascular disease, severe chronic liver disease, diabetes with end organ damage, chronic renal failure, and dementia.

for race, prices, ethnicity, age, sex, ZIP-code median income, and extensive diagnoses based on HCCs (Hierarchical Condition Categories); see Chandra et al. (2015) for details.[10] Finally, to allow comparison with other studies of regional variations, we adopt as our third measure of spending standard price-adjusted HRR Medicare expenditures from the Dartmouth Atlas.

Estimation issues. We estimate Equation (4) at the HRR level with one of the three spending measures described above as the dependent variable, and a variety of supply and demand variables, described below, on the right-hand side of the regression. Because of limited physician samples within HRRs, we limit our attention to the 74 HRRs with a minimum of three cardiologists (average = 5.6) and one primary care physician (average = 7.4) surveyed. This in turn leads to an attenuated sample of patients (N = 1,516), or an average of 20.5 respondents per HRR.[11]

We are concerned about measurement error of the constructed variables on the right-hand side of this regression specification. To address this concern, we bootstrap the entire estimation process 1,000 times to calculate all standard errors in HRR-level regressions, beginning with the construction of the HRR-level physician and patient samples, and proceeding through the HRR-level regression analysis.[12]

[10] The concern with HCCs is that because diagnoses are derived from billing data, aggressive physicians who admit patients to hospital and test for more disease tend to find more disease (Song et al., 2010; Finkelstein, et al., 2014). The time period does not exactly line up, but regions exhibit considerable temporal stability in health care spending patterns. Results are similar whether factors such as HCC scores and income are included as controls or not.

[11] In subsequent analysis that seeks to better explain who is a cowboy or comforter, we use the entire sample of valid physician measures (N=1533).

[12] Results with bootstrapped standard errors are presented in all tables and were not sensitive to a 10,000-replication bootstrap.

IV. Patient Survey and Vignettes

We use responses to 5 survey questions that ask patients about their likelihood of wanting unnecessary tests or cardiologist referrals in the case of new chest pain as well as preferences for comfort vs. intensive life-prolonging interventions in an end of life situation. The exact language used in these vignettes is reproduced in the first Appendix. Since the questions patients respond to are hypothetical and typically describe scenarios that have not yet happened, we think of them as x^D, or preferences not affected by physician advice.

Two of the questions in the patient survey relate to unnecessary care, asking respondents if they would like a test or cardiac referral even if their primary care physician did not think they needed one (Table 1).[13] Overall, 73 percent of patients wanted such a test and 56 percent wanted a cardiac referral. However, there is wide variation across regions in average responses to these questions. Figure 2 shows density plots of patient preferences for the main questions in the patient survey for the 74 HRRs considered (weighted by the number of patients per HRR). Superimposed on the Figure is the simulated distribution based on 1,000 bootstrap samples, with replacement, under the null hypothesis that individuals were randomly assigned to areas. P–values for over-dispersion are reported in the last column of Table 1; these indicate that the observed variation is significantly greater than can be explained by random variation.

Three other patient questions, grouped into two binary indicators, measure preferences for end-of-life care. One reflects patients' desire for aggressive care at the end of life: whether they would want to be put on a respirator if it would extend their life for either a week (one question) or a month (another question). The second question asked: if the patient reached a point at which they were feeling bad all of the time, would they want drugs to make them feel

[13] This question captures pure patient demand independent of what the physician wants. Note, however, that patients could still answer they would not seek an additional referral if they were unwilling to disagree with their physician.

better, even if those drugs might shorten their life. In each case, there is statistically significant variation across HRRs (Table 1).

Patients' preferences are sometimes correlated across questions. For example, the correlation coefficient between wanting an unneeded cardiac referral and wanting an unnecessary test is 0.44 ($p < .01$). But other comparisons point to very modest associations, for example a -0.29 correlation coefficient between wanting palliative care and wanting to be on a respirator at the end of life. In sensitivity analyses, we also considered using alternate measures of demand recovered from the patient survey. For example, we defined a patient with "high demand" for physician services as one who was quick to go to the doctor with a cough and was eager to have a chest x-ray even when a physician did not think it was necessary.

Potential biases in the patient survey. One concern with these questions is that they are hypothetical, and might not reflect true patient demand. However, Anthony et al. (2009) linked the survey respondents to the individuals' Medicare claims, and found a strong link between preferences and care-seeking behavior (e.g., physician visits) at the individual level (also see Mandelblatt et al., 2012). Questions about end-of-life care may appear abstract until the patient actually faces decisions related to end-of-life care, but we find strong regional patterns in responses to the questions, and the results are similar even with a subset of respondents age 80+ where end-of-life treatment decisions are more salient. Moreover, Medicare enrollees are increasingly choosing to create advanced directives that specify their end-of-life choices well before they face those decisions.[14]

V. Physician Vignettes

[14] According to one study, more than half of elderly Americans (and 29 percent of people aged 55-64) have advanced directives. The primary reason for not having an advanced directive was being unaware of their existence (Rao, et al., 2014).

The detailed clinical vignette questions used in the physician surveys are shown in the Appendix and summary statistics are presented in Table 1. With respect to the physician surveys, we begin by noting that prior research has established that "physicians do what they say they do" in a cardiac care setting. For example, physicians' self-reported testing intensity has been shown to be predictive of population-based rates of coronary angiography (Wennberg et al., 1997).

We first consider the vignette for Patient A, which asks how frequently the physician would schedule routine follow-up visits for patients with stable angina whose symptoms and cardiac risk factors are well controlled on current medical therapy (for cardiologists) or patients with hypertension (for primary care physicians). The response is unbounded, and expressed in months, but in practice individual physician responses ranged from 1 to 24 months. Figure 3a presents a HRR-level histogram of averages from the cardiology survey for all 74 HRRs studied.

How do these responses correspond to guidelines for managing chronic stable angina? The 2005 American College of Cardiology/American Heart Association [ACC/AHA] guidelines (Hunt et al., 2005) – what most cardiologists would have considered the "Bible" in cardiology at the time the survey was fielded – recommended follow-up every 4-12 months. However, even with these broad recommendations, we find that nearly one fifth (19 percent) of cardiologists in the sample recommend follow-up visits more frequently than every 4 months. These physicians were geographically clustered in a subset of HRRs ($p<.01$ in a test of the null of no geographic correlation) and the distribution of high follow-up cardiologists across HRRs is shown in Figure 3b.[15]

[15] The equivalent follow-up measure for primary care physicians is for a patient with well-controlled hypertension. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (U.S. Department of Health and Human Services, 2004), which would have been the most current guideline recommendation at the time, suggests follow-up for this type of patient every 3-6 months based on expert opinion.

We define a “high follow-up” physician as one who recommends follow-up visits more frequently than clinical guidelines would suggest, and a “low follow-up” physician as one who recommends follow-up visits less frequently than clinical guidelines would suggest. By this definition, roughly two percent of cardiologists and ten percent of PCPs in our data are classified as “low follow-up” physicians while 19 percent of cardiologists and 3 percent of PCPs in our data are classified as “high follow-up” physicians.[16]

The next two vignettes focus on patients with heart failure, a much more expensive setting. Heart failure is also a natural scenario to consider because it is common, the disease is chronic, prognosis is poor, and treatment is expensive. Vignettes for both Patients B and C ask questions about the treatment of Class IV heart failure, the most severe classification and one in which patients have symptoms at rest. In both scenarios the vignette patient is on maximal (presumably optimal) medications, and neither patient is a candidate for revascularization: Patient B has already had a coronary stent placed without symptom change, and Patient C is explicitly noted to not be a candidate for this procedure. The key differences between the two scenarios are patients’ ages (75 for Patient B, 85 for Patient C), the presence of asymptomatic non-sustained ventricular tachycardia in Patient B, and severe symptoms that resolve partially with increased oxygen in Patient C.

Cardiologists in the survey were asked about various interventions as well as palliative care for each of these patients. For Patient B, they were given five choices: three intensive treatments (repeat angiography; implantable cardiac defibrillator [ICD] placement, and pacemaker insertion), one involving medication (antiarrhythmic therapy), and palliative care.

[16] Office visits are not a large component of physicians’ incomes (or overall Medicare expenditures). Thus any correlation between the frequency of follow-up visits and overall expenditures would most likely be because frequent office visits are also associated with additional highly remunerated tests and interventions (such as echocardiography, stress imaging studies, and so forth) that further set in motion the “diagnostic-therapeutic cascade” (Lucas, et al., 2008).

Patient C also has three intensive options (admit to the ICU/CCU, placement of a coronary artery catheter, and pacemaker insertion), two less aggressive options (admit to the hospital [but not the ICU/CCU] for diuresis, and send home on increased oxygen and diuretics), and palliative care. In each case, cardiologists separately reported their likelihood of recommending each intervention on a 5-interval range from "always / almost always" to "never." Each response was independent of other responses, so for example, physicians could "frequently" recommend both palliative care and an intensive intervention.

We start with the obvious: regardless of the religious, political or moral persuasion of the physician, both men deserve a frank conversation about their prognosis and an ascertainment of their preferences for end-of-life care. The one-year mortality for either man exceeds 50 percent (Ahmed et al., 2006; Figure 1). If compliant with the guidelines, therefore, every one of the cardiologists should have answered "always/almost always", or at least "most of the time," to initiating or continuing discussions about palliative care.[17]

For Patient B, only 29 percent of cardiologists responded that they would initiate or continue discussions about palliative care "most of the time" or "always/almost always." For Patient C, 44 percent of cardiologists and 47 percent of primary care physicians were likely to recommend this course of action "most of the time" or "always/almost always." In both cases, physicians' recommendations fall far short of clinical guidelines. We classify the doctor as a "comforter" if the physician would discuss palliative care with the patient "always / almost always" or "most of the time" for both Patients B and C (among cardiologists) or for Patient C only (among primary care physicians, who did not have Patient B's vignette in their survey). In

[17] According to the AHA-ACC directives, "Patient and family education about options for formulating and implementing advance directives and the role of palliative and hospice care services with reevaluation for changing clinical status is recommended for patients with HF [heart failure] at the end of life." (Hunt, et al., 2005, p. e206)

our final sample, 27 percent of cardiologists and 47 percent of primary care physicians met this definition of a comforter.

We now turn to more controversial aspects of patient management. The language in the vignettes was carefully constructed to relate to the contemporaneous clinical guidelines. Several key aspects of Patient B rule out both the ICD and pacemaker insertion[18] and indeed the ACC-AHA guidelines explicitly recommend against the use of an ICD for Class IV patients potentially near death (Hunt et al., 2005; p. e206). On the other hand, both treatments are highly reimbursed.

Since Patient C is already on maximal medications and is not a candidate for revascularization, the physician's management goal should be to keep him as comfortable as possible. This should be accomplished in the least invasive manner possible (e.g., at home), and if that is not possible, in an uncomplicated setting, for example during admission to the hospital for simple diuresis. According to the ACC/AHA guidelines, no additional interventions are appropriate for this patient.[19] In fact, even a "simple" but invasive test, the pulmonary artery catheter, has been found to be of no marginal value over good clinical decision making in managing patients with CHF, and could even cause harm (ESCAPE, 2005).

Despite these guideline recommendations, physicians in our data show a great deal of enthusiasm for additional interventions. For Patient B, nearly one-third of the cardiologists surveyed (28 percent) would recommend a repeat angiography at least as frequently as "some of the time." Similarly, 62 percent of cardiologists recommend an ICD "most of the time," or "always/almost always," while 45 percent recommend a pacemaker at least "most of the time".

[18] This includes his advanced stage, his severe (Class IV) medication refractory heart failure, and the asymptomatic non-sustained nature of the ventricular tachycardia.

[19] Clinical improvement with a simple intervention (increasing his oxygen) also argues against more intensive interventions.

For Patient C, 18 percent recommend an ICU/CCU admission, 2 percent recommend a pulmonary artery catheter and 14 percent recommend a pacemaker at least “most of the time.”

Cardiologists’ responses on aggressive interventions are highly correlated across Patients B and C. Of the 28 percent (N=165) of cardiologists in the sample who would “frequently” or “always/almost always” recommend at least one of the high-intensity procedures for Patient C, 91 percent (N=150) would also frequently or always/almost always recommend at least one high-intensity intervention for Patient B. We use this overlap to define a “cowboy” cardiologist as a cardiologist who recommends at least one of the three possible intensive treatments for both Patients B and C “most of the time” or “always/almost always.” Because Vignette B was not presented to the primary care physicians, we use only their response to Vignette C to categorize them using the same criteria. In total, 25 percent of the cardiologists in our sample are classified as cowboys, as are 22 percent of primary care physicians.

All told, we test four measures of Z^S: high or low frequency of follow-up visits, a dummy variable for being a cowboy, and a dummy variable for being a comforter. How are these measures related? Table 2 shows that physicians with a low follow-up frequency are more likely to be comforters and less likely to be cowboys than physicians with a high follow-up frequency. Similarly, cowboy physicians are far less likely to be comforter physicians (even though doctors could be classified as both). Most differences are statistically significant.

Potential biases in physician vignettes. For the physician vignettes to be valid measures of physician beliefs, $\Psi c/M$, we need two conditions to be met. First, we want to rule out the possibility that $\Delta\alpha_i$ reflects real expertise, so that some of our physicians *should* be prescribing intensive treatments because of greater skill or expertise, as in Chandra and Staiger’s (2007) study of angioplasty (and thrombolytics) among AMI patients in the mid-1990s. In the 1996

American College of Cardiology / American Heart Association clinical guidelines (Ryan, et al., 1996) angioplasty was classified as a Category I treatment generally agreed to be "beneficial, useful, and effective", with a further recommendation that it be performed only in a "*timely fashion by individuals skilled in the procedure* and *supported by experienced personnel in high-volume centers.*" (Italics in the original). That Chandra and Staiger (2007) found variation in expertise across hospitals is entirely consistent with these guidelines.

By contrast, the use of ICDs for Patient B, recommended at least "most of the time" by 62 percent of cardiologists in the sample, is a Category III treatment, generally agreed to be "not useful/effective and in some cases may be harmful." (Gregoratos, et al., 2002). In other words, there is no evidence that this treatment (and most of the others) is effective or useful, for *any* physician. Support for this view comes from the fact that cowboy physicians live in regions where the HRR-level composite AMI Hospital Compare quality score is lower (Dartmouth Atlas, 2013), and are less likely to be board-certified, or practice in an academic setting.

We further considered the possibility that aggressiveness of end-of-life care may be related to overall quality. We analyzed the range of quality measures from the CMS Hospital Compare data (December 2005 data release[20], averaged at the HRR level), but found few additional statistically significant correlations between the fraction of aggressive physicians (cowboys) and quality measures. Of the seventeen quality measures considered, only two[21] had a statistically significant correlation with the fraction of cowboys in a region, however, for the vast majority of measures (88 percent), there was no correlation with the local fraction of cowboy physicians.

[20] Data available at: https://data.medicare.gov/data/archives/hospital-compare

[21] Those two quality measures that were negatively associated with the fraction of cowboys in an HRR were 1) AMI patients with left ventricular systolic dysfunction (and without contraindications) who were prescribed an ACE inhibitor or an ARB at hospital discharge and 2) Pneumonia patients who receive their first dose of antibiotic within 4 hours of arrival at the hospital.

The second potential concern with these estimates is that physicians in regions with sicker and more expensive patients could "fill in" missing characteristics of their patients, and opt for the more intensive procedures. If our measures of spending are imperfectly risk-adjusted, this could lead to a spurious correlation between physician beliefs and overall spending. Yet there is little evidence that physicians do "fill in" characteristics of patients; in one study, physicians treated actors who differed by race, gender, and social class, but presented with identically scripted descriptions of symptoms consistent with congestive heart failure (Arber, et al., 2006). While women were asked fewer questions and had fewer diagnostic tests ordered, there was no differences in treatment choices by race or social class.[22] Further, even if "fill-in" effects were observed, a patient's lower socioeconomic status or minority race has not been found to be predictive of aggressive care (Rao et al., 2003). In sum, (a) we do not find evidence in the clinical literature that "fill in" effects occur in the use of vignettes, (b) physicians are no more likely to recommend aggressive treatments for lower income or minority patients, and (c) even if physicians did recommend more aggressive treatments for minority Class IV CHF patients, there is no evidence from ACC/AHA clinical guidelines that they should. Still, in robustness tests, we also consider outcome measures that have already adjusted for income differences across patients.

VI. Summary Statistics and Model Estimates

We now proceed with our estimates of the models presented above. We first consider Equation (4), the relationship between area-level spending and local patient and physician

[22] Since our vignettes are for males, we are less worried about the gender differences found in this study. Also see Haider, et al. (2011) who found that medical student responses to clinical vignettes were not sensitive to patient race or occupation.

preferences. We then turn to Equation (6), modeling the factors leading physicians to be more and less aggressive.

Do Survey Responses Predict Regional Medicare Expenditures? We start with the basic relationship between area spending, patient preferences and physician preferences for the 74 HRRs with at least three cardiologists responses as well as at least one primary care physician response. Figure 4 shows scatter plots of area-level end of life spending vs. our measures of supply and demand for care. The measures we include here are the fraction of all physicians in the area who are cowboys (panel a), the fraction of physicians who are comforters (panel b), the fraction of physicians who recommend follow-up more frequently than recommended guidelines (panel c), and the share of patients who desire more aggressive care at the end of life (panel d). Each circle represents one HRR, and its size is proportional to the physician survey sample size for that HRR.

In the case of the three supply-side variables in Figure 4, the results are consistent with the theory: despite the relatively small sample sizes of physicians in each HRR, end of life spending is positively related to the cowboy ratio, negatively related to the comforter ratio, and positively related to the fraction of doctors recommending follow-up visits more frequently than clinical guidelines indicate. The demand variable, in contrast, is not strongly related to spending: the data points form more of a cloud than a line.

Table 3 explores these results more formally with regression estimates of the natural log of end-of-life expenditures, weighted by the number of physician observations per HRR and including controls for the fraction of PCPs among our survey responders. The standard errors presented in all HRR-level tables are calculated from 1000 bootstrap samples of patient and physician surveys. As the first column shows, the local proportion of cowboys and comforters predicts 36 percent of the observed regional variation in risk-adjusted end-of-life spending. Further, the estimated magnitudes are large: increasing the percentage of cowboys by one standard deviation (15 percentage points) is associated with a 12 percent increase in end-of-life expenditures, while increasing the fraction of comforters by one standard deviation (18 percentage points) implies a 5.8 percent reduction in expenditures. This relationship between spending and the local fractions of cowboys and comforters is robust to a number of alternative specifications (see the second Appendix for all results). First, the relationship holds when both cardiologists and primary care physicians are analyzed separately. Further, the observed patterns are robust to the exclusion of "small cells" (HRRs with fewer than 8 physician surveys) and controlling for percent poverty at the HRR level.

Column 2 of Table 3 shows that the indicator for high frequency follow-up recommendations is also a meaningful predictor of HRR-level end-of-life spending: conditional on the fraction of cowboys and comforters, an increase of one standard deviation (12 percentage points) of physicians who prefer to see patients more frequently than guidelines recommend is predicted to increase end-of-life spending by 13 percent. While the low frequency follow-up coefficient is meaningful in magnitude – implying roughly a 3 percent reduction in spending for a one standard deviation increase in low follow-up physicians – it is not statistically significant. Indeed, the combination of just these four variables about supplier beliefs alone can explain over 60 percent of the observed end-of-life spending variation in the 74 sample HRRs.[23]

The next two columns add measures of patient preferences to the regressions: the share of patients wishing to have unneeded tests, the share wanting to see an unneeded cardiologist, the

[23] As Black, et al. (2000) note, the OLS estimate is a lower bound and under weak assumptions, the expected value of the OLS parameter estimate is of smaller magnitude than the true parameter. (The R^2 is also a lower bound owing to measurement error.)

share preferring aggressive end-of-life care, and the share preferring comfortable end-of-life care. None of these variables are statistically significant at the 5 percent level. Even excluding the physician belief variables entirely, as in column 6, the R^2 from a regression including the patient preference variables is just 0.08. In further sensitivity analyses, we also considered using alternate and/or additional measures of patient demand recovered from the patient survey and these did not increase the explanatory power of our regression models, nor were the coefficients statistically significant at conventional levels.[24]

It is also possible that there is an interaction effect between patient preferences and physician beliefs – for example if aggressive physicians interact with patients with preferences for aggressive care to generate even more utilization (or conversely for comforter physicians and patients who demand palliative care). These hypotheses are considered in Table 4. Column 1 of the table repeats column 5 of Table 3 for reference. The subsequent columns add interaction terms. As seen in column 2, however, there is little consistent evidence for the interactive aggressiveness hypothesis; the interaction between cowboy physicians and patients with aggressive preferences is negative (not positive as the above theory would suggest), and while the coefficient between comforter physicians and patients is negative (column 3), it is not statistically significant.

Column 4 of Table 4 repeats the analysis in column 1, but uses the log of average Medicare expenditures for AMI patients in the year after admission to the hospital. As noted above, this measure adjusts most carefully for differences across regions in underlying health

[24] In sensitivity analyses, we also defined a patient with "high demand" for physician services as one who was quick to go to the doctor with a cough and was eager to have a chest x-ray even when a physician did not think it was necessary. Results were similar. Nor were our results driven by differences between east-coast versus west-coast medical practice; similar results were obtained when east and west coasts were estimated separately.

status and poverty rates that might affect spending. The estimated coefficients suggest relationships similar to those in column 1, but the coefficients are smaller in magnitude as is the R^2 value (0.34 versus 0.64).

Column 5 of Table 4 repeats the analyses in column 1, but uses total average per beneficiary Medicare expenditures (adjusted for prices, age, sex, and race/ethnicity) as the dependent variable. In the combined sample, the fraction of cowboys in an HRR is a consistently strong predictor of spending across models. Moreover, although R^2 values are smaller in these models, supply-side factors continue to explain more of the variation in spending than demand-side factors.

How much would Medicare expenditures change in a counterfactual setting in which there were no cowboys, all physicians were comforters, and all physicians met guidelines for follow-up care? While we recognize the conjectural nature of this calculation, setting the fraction of cowboys to 0 and comforters and guideline consistency to 1 implies a decline of 35 percent in end-of-life expenditures and a decline of 12 percent in total Medicare expenditures. Extending this estimate to the entire health care economy suggests that roughly 2 percent of GDP in health care spending is associated with treatments without clinical evidence. These comparisons point to the importance of physician beliefs in explaining regional (and national) utilization patterns.

What factors predict physician responses to the vignettes? To this point, we have shown that aggregated physician vignette responses explain regional patterns in spending, and that these responses vary across areas far more than would be expected given random variation. The obvious question is then: what explains this variation? In this section, we estimate the model in Equation (6), using the entire sample of physicians to test for the relative importance of financial and organizational factors in explaining physician recommendations. We interpret the residual, conditional on financial and organizational factors, as pure physician "beliefs."

Various measures of financial circumstances are reported in Table 1 for all physicians. We focus on three: the share of patients for whom they are reimbursed on a capitated basis (on average, 16 percent), the share of patients whose primary health insurance is Medicare (43 percent), and the share of a physician's patients on Medicaid (10 percent), with capitated payment and Medicaid generally associated with lower marginal reimbursement.[25]

A second set of questions asks about characteristics of the physician and her practice. Twenty-nine percent work in small practices (solo or 2-person), 58 percent work in single or multi-specialty group practices, and 13 percent work in HMOs or hospital-based practices. We also observe a number of characteristics about the physician, including age, gender, whether she is board certified, and the number of days per week she spends seeing patients.

Third, the survey asks about a physician's self-assessed responsiveness to external incentives over the past year, including how frequently, if ever, in the past 12 months she had intervened on a patient for non-clinical reasons. We create a set of binary variables that indicates whether a physician responded to each set of incentives at least "sometimes" (i.e. "sometimes" or "frequently") over the past year. Twenty-eight percent of cardiologists reported that they had sometimes or frequently performed a cardiac catheterization because of the expectations of the referring physician and 40 percent of all physicians reported having intervened medically because of a colleague's expectations (Table 1).

We next consider regression analysis. Table 5 presents coefficients from a linear probability model with HRR-level random effects for three regressions at the physician level.

---

[25] The survey also asks about the fraction of Medicare patients each physician sees. All analyses below are robust to including this fraction among the independent variables, but it was always statistically insignificant and its exclusion did not affect other coefficients.

Our dependent variables are binary indicators for whether the physician is a cowboy (column 1), a comforter (column 2), or recommends high frequency follow-up (column 3). In each model, we include basic physician demographics: age, gender, board certification status, whether the physician is a cardiologist, days per week spent seeing patients, as well as cardiologists per 100,000 Medicare beneficiaries.

Notably, some physician characteristics are predictive of physician types: male physicians in the sample are somewhat more likely to be cowboys and less likely to be comforters than female doctors. Older physicians are more likely to be both cowboys and high follow-up: at the mean age of 52 years, a 1 standard deviation increase in physician age (9.9 years) is associated with a 4.0 percent increase in probability of being a cowboy and a similar 4.0 percent increase in probability of being a high follow-up physician. Board certification – a rough marker for physician quality – is negatively associated with cowboy status and high follow-up frequency, but not statistically significantly so.

The substitution effect implies that lower incremental reimbursements associated with Medicaid and capitated patients would lead to fewer interventions and more palliative care. Table 5 shows that physicians with a larger fraction of Medicaid and (but not capitated patients) are more likely to be cowboys and high-follow-up physicians, rejecting the dominance of the substitution effect. One may appeal again to a dominant income effect to explain these patterns.

Some organizational factors are strongly associated with physician beliefs about appropriate practice. Physicians in solo or 2-person practices are far more likely to be both cowboys and high follow-up doctors than physicians in single or multi-specialty group practices or physicians who are part of an HMO or a hospital-based practice. Whether cardiologists accommodate referring physicians – a financial factor (since cardiologists benefit financially from future referrals) as well as an organizational one – is a positive predictor of being a cowboy among cardiologists.[26] Finally, malpractice concerns are predictive of neither cowboy nor comforter status.[27]

The explanatory power of these regressions is quite modest – between 4 and 14 percent – suggesting that a considerable degree of the remaining variation is the consequence of physician *beliefs* regarding the productivity of treatments, rather than behaviors systematically related to financial, organizational, or other factors.

As a final exercise, we include these financial, organizational, and responsiveness variables, aggregated up to the HRR level, in a regression that seeks to explain the variation in log end-of-life spending – an expanded counterpart to Table 4. These results are presented in the Appendix. None of the additional variables are statistically significant, nor do they add appreciably to the explanatory power of the regression. Physician beliefs, independent of financial or organizational factors, appear to explain a great deal of why physicians are cowboys or comforters and how the frequencies of these typologies, in turn, are related to overall spending.

VII. Conclusion and Implications

While there is a good deal of regional variation in medical spending and care utilization in the U.S. and elsewhere, there is little agreement about the causes of such variations. Do they

[26] Note that the question on responding to referring doctor expectations appeared in the cardiologist survey only, and so reflects the preferences of cardiologists only; the cardiology dummy variable therefore reflects both the pure effect of being a practicing cardiologist, and a secondary adjustment arising from the referral question being set to zero for all primary care physicians.

[27] Perhaps because procedures performed on high-risk patients (such as Patients B and C) can increase the risk of a malpractice suit (Currie and MacLeod, 2008).

arise from variation in patient demand, from variation in physician behavior, or both? In this paper, we find that regional measures of patient demand as measured by responses to a nationwide survey have only modest predictive association with regional end-of-life expenditures (Anthony, et al., 2009; Baker et al., 2014). By contrast, regionally aggregated measures from physician vignette responses regarding treatment options can explain a substantial degree of observed regional variation in utilization in the U.S. Medicare population even after adjusting for demand. Generally, prior studies inferred practice variations as the residual from an area model, leading to estimates being biased either upward (because of unobserved regional factors) or downward (because of flawed risk-adjustment practices, as in Song et al., 2010). In contrast, this paper uses physician vignettes to explain regional variation. Indeed, it seems that simply presenting a dozen randomly chosen physicians in a region with several vignettes would allow an observer to form a remarkably accurate prediction of actual regional Medicare expenditure rates.

A back-of-the-envelope calculation using our regression results implies that, were all physicians in the 74 HRRs studied to follow professional guidelines, end-of-life Medicare expenditures in these regions would be expected to be 35 percent lower and overall Medicare expenditures 12 percent lower. To the extent that incremental treatments by "cowboys" (or those received from too-frequent revisits) yield little or no benefit to patients, our results are consistent with roughly 2 percent of GDP devoted to treatments that are not supported by clinical evidence.

We then turned to the factors that lead physicians to have different preferences. We find that the traditional factors in supplier-induced demand models, such as the fraction of patients paid through capitation (or on Medicaid), or physicians' responsiveness to financial factors, play a relatively small role in explaining equilibrium variations in utilization patterns. Organizational factors, such as accommodating colleagues, help to explain only a small amount of observed variation in individual intervention decisions. Instead, differences in physician beliefs about the effectiveness of treatments explain the lion's share of inter-regional variation in Medicare expenditures.[28]

Unfortunately, the data we consider in this study cannot shed light on how these differences in physician beliefs arise. Previous hypotheses have included variation in medical training (Epstein and Nicholson, 2009) or their personal experiences with different interventions (Levine-Taub, et al., 2015). Simple heterogeneity in physician beliefs cannot explain regional variation in expenditures, since the observed regional patterns in physician beliefs exhibit far greater inter-region variation than would be expected due to chance alone. Rather, spatial correlation in beliefs is required in order to explain the regional patterns we see. We do find that physicians' propensity to intervene for non-clinical reasons is related to the expectations of physicians with whom they regularly interact, a result consistent with network models. Similarly, Molitor (2011) finds that cardiologists who move to more or less aggressive regions change their practice style to better conform to local norms. However, we are still left with questions as to how and why some regions become more aggressive than others.

Our results do not imply that economic incentives are unimportant. Clearly, changes in payment margins have a large impact on behavior, as has been shown in a variety of settings. But the prevalence of geographic variations in European countries, where economic incentives are often nearly entirely blunted (Skinner, 2012), is also consistent with the view that physician

[28] This result is consistent with Epstein and Nicholson (2009), who find large variations in Cesarean section surgical rates among obstetricians within the same practice, and with Chassin's (1993) "Enthusiasm Hypothesis" – that regional differences in the use of health care services are caused by differences in the prevalence of physicians who are enthusiasts for those services.

beliefs play a large role in explaining such variations. A better understanding of both how physician beliefs form, and how they can be shaped, is a key challenge for future research.

References

Ahmed Ali, Wilbert S Aronow, and Jerome L Fleg, 2006. "Higher New York Heart Association Classes and Increased Mortality and Hospitalization in Heart Failure Patients with Preserved Left Ventricular Function," *American Heart Journal*, 151(2): 444-450.

Ameriks, John, Andrew Caplin, Steven Laufer, and Stijn van Nieuwerburgh, 2011. "The Joy of Giving or Assisted Living? Using Strategic Surveys to Separate Public Care Aversion from Bequest Motives," *Journal of Finance*, 66(2) (April), 519-561.

Anthony, Denise L., M. Brooke Herndon, Patricia M. Gallagher, Amber E. Barnato, Julie PW Bynum, Daniel J. Gottlieb, Elliott S. Fisher, and Jonathan S. Skinner. "How much do patients' preferences contribute to resource use?."*Health Affairs* 28, no. 3 (2009): 864-873.

Arber, Sara, John McKinlay, Ann Adams, Lisa Marceau, Carol Link, and Amy O'Donnell. "Patient characteristics and inequalities in doctors' diagnostic and management strategies relating to CHD: a video-simulation experiment." *Social science & medicine* 62, no. 1 (2006): 103-115.

Baker, Laurence C., M. Kate Bundorf, and Daniel P. Kessler, 2014. "Patients' Preferences Explain a Small but Significant Share of Regional Variation in Medicare Spending," *Health Affairs* 33(6), pp. 957-63.

Barnato, Amber E., M. Brooke Herndon, Denise L. Anthony, Patricia M. Gallagher, Jonathan S. Skinner, Julie PW Bynum, and Elliott S. Fisher. "Are regional variations in end-of-life care intensity explained by patient preferences?: A Study of the US Medicare Population." *Medical care* 45, no. 5 (2007): 386.

Baumann, Andrea O., Raisa B. Deber, Gail G. Thompson, 1991, "Overconfidence Among Physicians and Nurses: The 'Micro-Certainty, Macro-Uncertainty' Phenomenon," *Social Science and Medicine* 32(1): 167-74.

Black, Dan A.; Mark C. Berger; Frank A. Scott, 2000. "Bounding Parameter Estimates with Nonclassical Measurement Error," *Journal of the American Statistical Association*, Vol. 95(451), pp. 739-748.

Campbell, Eric G., Sowmya R. Rao, Catherine M. DesRoches, Lisa I. Iezzoni, Christine Vogeli, Dragana Bolcic-Jankovic, and Paola D. Miralles, "Physician Professionalism and Changes in Physician-Industry Relationships From 2004 to 2009," *Archives of Internal Medicine*, 170(20), pp. 1820-26.

Chandra, Amitabh, and Douglas O. Staiger. 2007. "Productivity Spillovers in Healthcare: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy* 115, 103-140.

Chandra, Amitabh, and Jonathan Skinner, 2012, "Productivity growth and expenditure growth in U.S. Health care," *Journal of Economic Literature* 50(3) (September): 645-80.

Chandra, Amitabh, Dhruv Khullar, and Thomas H. Lee. "Addressing the Challenge of Gray-Zone Medicine." *New England Journal of Medicine* 372, no. 3 (2015): 203-205.

Chassin, Mark R. "Explaining geographic variations: the enthusiasm hypothesis." Medical care (1993): YS37-YS44.

Clemens, Jeffrey, and Joshua D. Gottlieb. "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?." *The American economic review*104, no. 4 (2014): 1320.

Currie, Janet, and W. Bentley MacLeod, 2008. "First Do No Harm? Tort Reform and Birth Outcomes," *Quarterly Journal of Economics* 123(2): 795-830.

Dartmouth Atlas, 2013. http://www.dartmouthatlas.org/tools/downloads.aspx

Doyle Jr, Joseph J. "Returns to local-area healthcare spending: evidence from health shocks to patients far from home." *American economic journal. Applied economics* 3, no. 3 (2011): 221.

Dranove, David, and Paul Wehner, 1994. "Physician-induced demand for childbirths*," Journal of Health Economics* 13(1) (March): 61–73.

Epstein, Andrew J., and Sean Nicholson, 2009. "The formation and evolution of physician treatment styles: An application to cesarean sections," *Journal of Health Economics,* 28(6) (December): 1126-1140.

(ESCAPE) Evaluation Study of Congestive Heart Failure and Pulmonary Artery Catheterization Effectiveness: The ESCAPE Trial. *JAMA.* 2005; 294(13): 1625-1633.

Evans, Spencer C., Michael C. Roberts, Jared W. Keeley, Jennifer B. Blossom, Christina M. Amaro, et al., 2015. "Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies. *International Journal of Clinical and Health Psychology*, 2015 (pre-print).

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. *Sources of geographic variation in health care: Evidence from patient migration*. No. w20789. National Bureau of Economic Research, 2014.

Fisher, Elliott S., David E. Wennberg, Terese A. Stukel, Daniel J. Gottlieb, Francis Lee Lucas, et al., 2003, "The implications of regional variations in Medicare spending. Part 2: Health outcomes and satisfaction with care," *Ann Intern Med* 138, Feb 18, 288-98.

Gogineni, Keerthi, Katherine L. Shuman, Derek Chinn, Nicole B. Gabler, and Ezekiel J. Emanuel. "Patient Demands and Requests for Cancer Tests and Treatments." *JAMA Oncology* 1, no. 1 (2015): 33-39.

Gregoratos, Gabriel, Jonathan Abrams, Andrew E. Epstein, Roger A. Freedman, David L. Hayes, Mark A. Hlatky, Richard E. Kerber et al. "ACC/AHA/NASPE 2002 guideline update for implantation of cardiac pacemakers and antiarrhythmia devices—summary article: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (ACC/AHA/NASPE Committee to Update the 1998 Pacemaker Guidelines)." *Journal of the American College of Cardiology* 40, no. 9 (2002): 1703-1719.

Grytten, Jostein, and Rune Sørensen. "Practice variation and physician-specific effects." *Journal of health economics* 22, no. 3 (2003): 403-418.

Haider, Adil H., Janel Sexton, N. Sriram, Lisa A. Cooper, David T. Efron, Sandra Swoboda, Cassandra V. Villegas et al. "Association of unconscious race and social class bias with vignette-based clinical assessments by medical students." JAMA 306, no. 9 (2011): 942-951.

Hunt, Sharon Ann, William T. Abraham, Marshall H. Chin, Arthur M. Feldman, Gary S. Francis, et al., 2005. "ACC/AHA 2005 Guideline Update for the Diagnosis and Management of Chronic Heart Failure in the Adult," *Circulation*. 2005;112:e154-e235; (September 13).

Hunt, Sharon Ann, William T. Abraham, Marshall H. Chin, Arthur M. Feldman, Gary S. Francis, et al., 2009. "2009 Focused Update Incorporated Into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Chronic Heart Failure in the Adult," *Circulation* 119:e391-e479 (March 26).

Institute of Medicine, Committee on Geographic Variation in Health Care Spending and Promotion of High-Value Care; Board on Health Care Services; Newhouse JP, Garber AM, Graham RP, et al., editors. Variation in Health Care Spending: Target Decision Making, Not Geography. Washington (DC): National Academies Press (US); 2013 Oct 1. Available from: http://www.ncbi.nlm.nih.gov/books/NBK201647/

Jacobson, Mireille, A. James O'Malley, Craig C. Earle, Juliana Pakes, Peter Gaccione and Joseph P. Newhouse, 2006. "Does Reimbursement Influence Chemotherapy Treatment For Cancer Patients?" *Health Affairs*, 25(2): 437-443.

Kelley, Amy S., Susan L. Ettner, R. Sean Morrison, Qingling Du, Neil S. Wenger, and Catherine A. Sarkisian. "Determinants of medical expenditures in the last 6 months of life." *Annals of internal medicine* 154, no. 4 (2011): 235-242.

Levine-Taub, Anna A., Anton Kolotilin, Robert S. Gibbons, Ernst R. Berndt, 2015. "The Diversity of Concentrated Prescribing Behavior: An Application to Antipsychotics," *Journal of Health Economics*, 40: 26-39.

Lucas, Francis Lee, Andrea E. Siewers, David J. Malenka, and David E. Wennberg, 2008. "Diagnostic-therapeutic cascade revisited. Coronary angiography, coronary artery bypass surgery, and percutaneous coronary intervention in the modern era." *Circulation*, 118: 2797-2802.

Lucas, Frances Lee, Brenda E. Sirovich, Patricia M. Gallagher, Andrea E. Siewers, and David E. Wennberg, 2010. “Variation in Cardiologists’ Propensity to Test and Treat: Is it Associated with Regional Variation in Utilization?” *Circulation : Cardiovascular Quality and Outcomes* 3 : 253-60.

Mandelblatt Jeanne S., Leigh Anne Faul, Gheorghe Luta, Solomon B. Makgoeng, Claudine Isaacs, et al., 2012. “Patient and physician decision styles and breast cancer chemotherapy use in older women: Cancer and Leukemia Group B protocol 369901” *Journal of Clinical Oncology* July 20;30(21): 2609-14.

McGuire, Thomas G., 2011, Physician agency and payment for primary medical care, in Sherry Glied, and Peter C. Smith, eds.: *The Oxford handbook of health economics* (Oxford University Press).

McGuire, Thomas G., Pauly, M., 1991. “Physician response to fee changes with multiple payers.” *Journal of Health Economics*, 10 (4), 385-410.

Moliter, David, “The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration,” MIT, 2011. http://economics.mit.edu/files/7301.

Pauly, Mark, 1980. *Doctors and their workshops: Economic models of physician behavior* (University of Chicago Press, Chicago).

Peabody John W., Jeff Luck, Peter Glassman, Sharad Jain, Joyce Hansen, Maureen Spell, and Martin Lee, 2004. “Measuring the quality of physician practice by using clinical vignettes: a prospective validation study,” *Annals of Internal Medicine* Nov 16;141(10):771-80.

Peabody, John W., Jeff Luck, Peter Glassman, Timothy R. Dresselhaus, and Martin Lee. "Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality."*JAMA* 283, no. 13 (2000): 1715-1722.

Rao, Sunil V., Padma Kaul, L.Kristin Newby, A.Michael Lincoff, Judith Hochman, et al., 2003. “Poverty, Process of Care, and Outcome in Acute Coronary Syndromes,” *Journal of the American College of Cardiology* 41(11): 1948-54.

Rao, Jaya K., Lynda A. Anderson, Feng-Chang Lin, and Jeffrey P. Laux, 2014. “Completion of Advance Directives among U.S. Consumers,” *American Journal of Preventive Medicine* 46(1): 65-70.

Ransohoff, David F., Mary McNaughton Collins, and Floyd J. Fowler Jr., 2002. “Why is prostate cancer screening so common when the evidence is so uncertain? A system without negative feedback,” *The American Journal of Medicine* 113(8): 663-67.

Ryan, Thomas J., Jeffrey L. Anderson, Elliott M. Antman, Blaine A. Braniff, Neil H. Brooks, et al., 1996, “ACC/AHA Guidelines for the Management of Patients With Acute Myocardial Infarction: A Report of the American College of Cardiology/American Heart Association Task

Force on Practice Guidelines (Committee on Management of Acute Myocardial Infarction)," *Circulation* 94:2341-50.

Sirovich, Brenda, Patricia M. Gallagher, David E. Wennberg, and Elliott S. Fisher, 2008. "Discretionary Decision Making by Primary Care Physicians and the Cost of U.S. Health Care," *Health Affairs* 27, May-Jun, 813-23.

Skinner, Jonathan, 2012. "Causes and Consequences of Geographic Variation in Health Care" in Thomas G. McGuire, Mark Pauly, and Pedro P. Barros (eds.) *Handbook of Health Economics* Vol. 2, North Holland.

Song, Yunjie, Jonathan Skinner, Julie Bynum, Jason Sutherland, John E. Wennberg, and Elliott S. Fisher, 2010. "Regional Variations in Diagnostic Practices," *New England Journal of Medicine* 363(1): 45-53.

U.S. Department of Health and Human Services, 2004. Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC7). NIH Publication No. 04-5230.

Wennberg, John E., Benjamin A. Barnes and Michael Zubkoff, 1982. "Professional Uncertainty and the Problem of Supplier-Induced Demand," *Social Science and Medicine* 16(7): 811–824.

Wennberg, David E., John D. Dickens, Jr., Lois Biener, Floyd J. Fowler, Jr., David N. Soule, and Robert B. Keller, 1997. "Do Physicians Do What they Say? The Inclination to Test and Its Association with Coronary Angiography Rates." *Journal of General Internal Medicine* 12(3):172-176.

Yasaitis, Laura C., Julie P. Bynum, and Jonathan Skinner, 2013. "Association between Physician Supply, Local Practice Norms, and Outpatient Visit Rates," *Medical Care* 51(6): 524-31.

Figure 1: Variations in Equilibrium: Differences in λ and Differences in Actual or Perceived Productivity

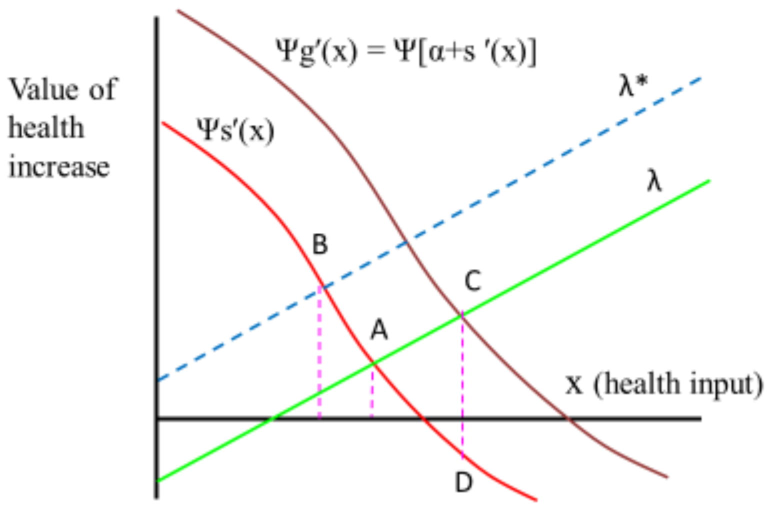

Figure 2: Distributions of Patient Preferences vs. Simulated Distributions (based on 1000 bootstrap samples with replacement)

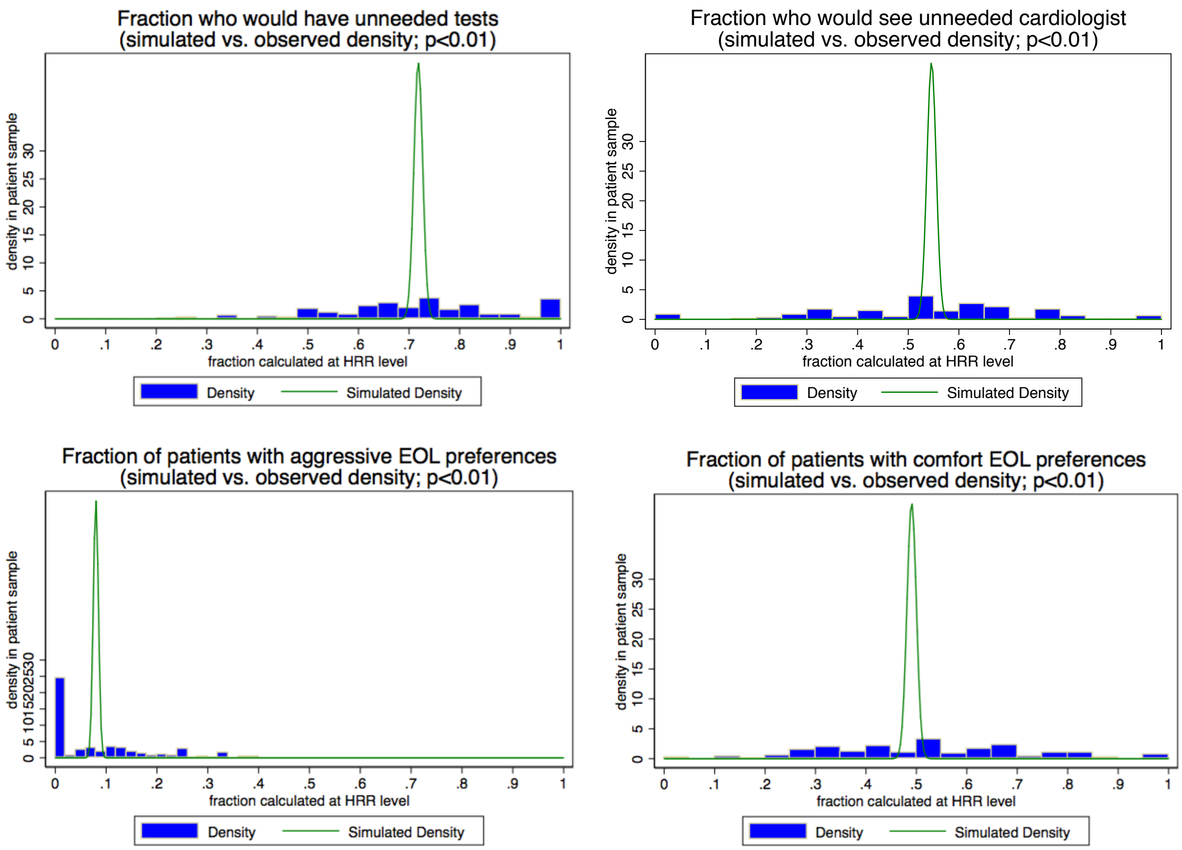

Figure 3a: Distribution of Length of Time before Next Visit for Patient with Well-Controlled Angina (Cardiologist HRR-Level Averages)

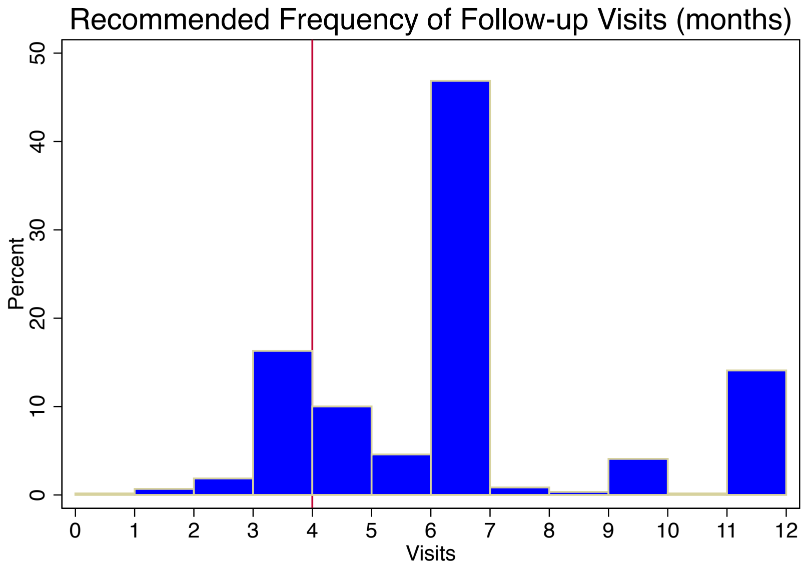

Figure 3b: Distribution of High Follow-Up Cardiologists and Geographic Correlation (HRR-Level Averages)

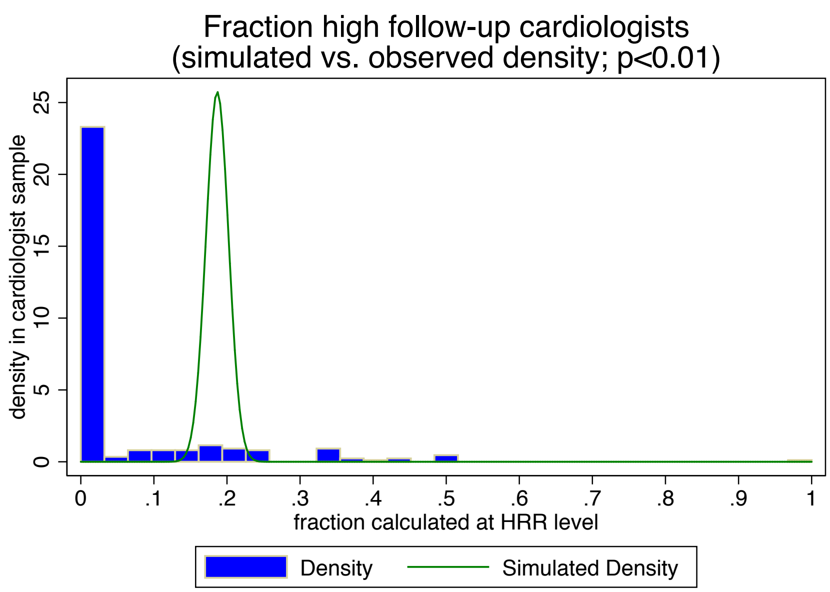

Figure 4: Ln of 2-year End-of-Life Regional Spending vs. Selected Independent Variables