Implications of Physician Ethics, Billing Norms, and Service Cost Structures for Medicare’s Fee Schedule

Clemens, Jeffrey

University of California at San Diego, National Bureau of Economic Research

2 February 2014

Online at https://mpra.ub.uni-muenchen.de/73392/
MPRA Paper No. 73392, posted 30 Aug 2016 06:35 UTC
Implications of Physician Ethics, Billing Norms, and Service Cost Structures for Medicare’s Fee Schedule

Jeffrey Clemens*
UCSD and NBER
February 3, 2014

Abstract
Medicare Part B pays physicians through a fixed fee schedule designed loosely as a system of average-cost reimbursement. This paper examines four difficulties faced by systems of this kind. First, Medicare’s payment model would be improved if it accounted for the medical value and cost-effectiveness of treatments in addition to their input costs. Second, uniformly applied fee schedules are inefficient when physicians vary in their approaches to medical practice. Allowing Medicare to account for regional differences in practice styles, which are substantial, may have significant benefits. Third, differences in physicians’ billing practices have similar, largely unstudied, implications. Proficient billers receive relatively high payments for incremental service provision, resulting in unintended variation in effective wages. Fourth, differences in services’ cost structures point to an additional weakness in Medicare Part B’s payment model. Average-cost reimbursement implies larger profit margins for capital-intensive services than for labor-intensive services. As implemented, Medicare’s fee schedule has encouraged significant expansions in the adoption, utilization, and development of capital-intensive tests and treatments.

*Clemens: jeffclemens@ucsd.edu
Most insurers compensate physicians through fee schedules. Determining these fees can be difficult, as insurers and physicians determine payments for thousands of distinct billing codes. The history of Medicare’s Resource Based Relative Value Scale (RBRVS), the basis for its outpatient fee schedule, illustrates the point. Its development and initial implementation spanned roughly a decade (Newhouse 2002). Private insurers’ adoption of RBRVS-based models points to both the cost of developing payment systems from scratch and the value of getting Medicare’s right.1

This paper analyzes four complications that arise in the context of Medicare’s fee schedule. First, Medicare’s payment model suffers from its exclusive focus on cost; it accounts for neither the health benefits nor the cost-effectiveness of care. Second, physicians vary in the norms by which they approach the practice of medicine (Cutler, Skinner, Stern and Wennberg 2013). Third, health care providers vary in their billing practices. This includes both their coding of patient conditions and the thoroughness with which they request payment for billable services (Dafny 2005). Finally, services differ significantly in their cost structures. Fee schedules that reimburse on an average-cost basis have very different implications for the profit margins associated with capital- and labor-intensive services.

I examine these complications through the lens of the classic Ellis and McGuire (1986) model of physician payment. Physician payment systems face a fundamental tradeoff. Paying on a per-unit, or fee-for-service, basis risks encouraging excess service provision; paying on a salary basis risks rewarding those who skimp on the quality and quantity of care. Ellis and McGuire characterize the optimal strength of per-unit incentives in light of these competing concerns. That is, they derive the mix of salary and per-unit payment that leads physicians to provide care to the point at which marginal benefits equal marginal costs. This optimal mix depends crucially on the weight physicians place on patient health, relative to income,

---

1 See Nandedkar (2011), Gesme and Wiseman (2010), and Fontes (2013) for practitioner characterizations of payment models and the process by which physicians negotiate with insurers.
when deciding how much care to provide.

As previously recognized, no single combination of salary and per-unit payment is optimal when physicians vary in their concern for patient health (Allard, Jelovac and Leger 2011, Godager 2013). I show that this conclusion applies to their billing norms as well. Both considerations imply that there may be value in increasing the flexibility with which Medicare’s fee schedule is implemented. Specifically, there may be gains from allowing claims administrators (Medicare’s regional “carriers”) to account for regional differences in physicians’ practice and billing norms. The costs associated with variation in billing and practice norms also have implications for physician training. Outcomes could be improved by training physicians more uniformly in these particular respects.

At the national level, adjusting payments for differences in the cost structures of capital- and labor-intensive services may also have significant value. Average-cost reimbursement implies larger profit margins for capital-intensive services than for labor-intensive services. This appears to have been exacerbated by Medicare’s initial estimates of capacity utilization for advanced imaging equipment, including MRI machines (Lowes 2013). Understating utilization led Medicare’s average-cost approach to overpay on a per-service basis. A growing body of research, discussed below, finds that the resulting incentives encouraged significant expansions in the adoption, utilization, and development of capital-intensive tests and treatments.2

The paper proceeds as follows. Section I develops a simplified version of Ellis and McGuire’s payment model. Section II emphasizes the relevance of physicians’ ethics with respect to the value they place on patient health. Section III assesses the importance of billing practices, which determine how physicians convert quantities of care provision into

---

reimbursable claims. Section IV explores the relevance of variation in the cost-structure of different services.

1 The Ellis and McGuire Model of Physician Payment

This section develops a simplified version of Ellis and McGuire’s model of physician payment. Ellis and McGuire answer the following question: what payment method will lead physicians to supply the socially desired quantity of care, namely the quantity such that care’s marginal cost equals its marginal benefit? Payments include a fixed, salary-like component, \( \alpha \), and a reimbursement per unit of care, \( r \). A physician providing \( q \) units of care thus has revenue of \( \text{Revenue} = \alpha + r \times q \). Letting the function \( c(q) \) describe the total cost of providing \( q \) units of care, net profit is \( \pi = \alpha + r \times q - c(q) \).

A key parameter describing physician behavior is the extent to which physicians value patient health relative to net income. If \( \theta \), bounded between 0 and 1, describes the weight placed on net income and \( 1 - \theta \) describes the weight placed on patient health, then the physician’s utility function is

\[
U = \theta[\alpha + r \times q - c(q)] + (1 - \theta)b(q). \tag{1}
\]

The \( b(q) \) term describes the net benefit patients obtain from receiving \( q \) units of care. When providing an incremental unit of care, physicians thus generate additional revenue of \( r \), incur marginal costs of \( MC(q) = c' \), and take some pleasure in their patients’ realization of the marginal benefits of care, \( MB(q) = b' \). The physician’s marginal utility from providing an additional unit of care is thus
\[ MU(q) = \frac{dU}{dq} = \theta[r - MC(q)] + (1 - \theta)MB(q). \]  

(2)

The physician maximizes his or her utility by choosing \( q \) such that

\[ \theta r + (1 - \theta)MB(q) = \theta MC(q). \]  

(3)

If marginal benefits are strictly decreasing and marginal costs are strictly increasing, the solution to (3) will be unique.

The manager of the insurance arrangement seeks to maximize the net value of the care provided to its beneficiaries. Knowing that physicians behave in the manner described above, the insurer arranges the physicians’ incentives to induce them to provide care until its marginal benefit equals its marginal cost. The optimal per-unit payment can thus be found by substituting \( MC(q^*) = MB(q^*) \) into equation (3). The algebra below yields an expression for the optimal payment, \( r^* \):

\[
\begin{align*}
\theta r^* &= \theta MB(q^*) - (1 - \theta)MB(q^*) \\
&= \theta MB(q^*) - MB(q^*) + \theta MB(q^*) \\
&= 2\theta MB(q^*) - MB(q^*) \\
&= MB(q^*)[2\theta - 1] \\
\end{align*}
\]

\[ \Rightarrow r^* = \frac{MB(q)[2\theta - 1]}{\theta} = MB(q^*)[2 - \frac{1}{\theta}]. \]  

(4)
The above expression for the optimal reimbursement rate has two implications. Intuitively, optimal reimbursement rates are high when the benefits from care are large. All else equal, reimbursements should be set to encourage the provision of large quantities of services that provide substantial health benefits. Notably, Medicare is legislatively bound to base payments on input costs. Consequently, it takes neither health benefits nor cost-effectiveness estimates into account.

Second, the optimal reimbursement rate is increasing in the weight physicians place on net income, and thus decreasing in the weight they place on patient benefits. This can be seen from the fact that $\theta$ appears in the denominator of a term that enters negatively on the right-hand side; as $\theta$ rises, $\frac{1}{\theta}$ declines and $-\frac{1}{\theta}$ rises. When physicians take pleasure in improving patient health, optimal payment schemes lean more heavily towards fixed salary payments. For low $\theta$ physicians, augmenting the agency benefit with strong financial incentives induces inefficiently large quantities of care.

2 Implications of Physician Agency

This section emphasizes two implications of variation in the value physicians place on patient health. As widely recognized but often left unsaid, the Ellis and McGuire framework shows that a single payment scheme can only result in optimal care provision when all physicians place the same weight on patient benefits. When the strength of physician agency varies, a given payment scheme will either lead some physicians to supply too much care or others to provide too little. When payments are tailored towards physicians who place low weight on patient benefits, for example, excess care will tend to be provided by those who highly value patient health. The former physicians require significant financial incentives to provide beneficial care while the latter require relatively small per-unit payments.

When payments are fixed through a national fee schedule, as in Medicare Part B, variation
in physician agency implies inefficiencies in care provision. This conclusion has two related implications. First, a large body of research describes substantial differences in the practice of medicine across regions of the United States (Wennberg, Fisher, Skinner et al. 2002).\(^3\) Adjusted for regional price differences and demographics, per-beneficiary Medicare spending ranges from $5,000 in Honolulu to $15,000 in Miami (Gottlieb, Zhou, Song, Andrews, Skinner and Sutherland 2010). Cutler et al. (2013) provide evidence that these variations are driven by differences in the way physicians approach medicine rather than differences in health care demand. Consequently, there may be significant value in allowing Medicare’s regional carriers, with which it contracts for the administration of Part B claims, to modify payments in response to regional practice styles.

Second, within a market or payment region, identifying the styles of individual practices and basing payments upon them would be difficult if not impossible (Allard et al. 2011). Some degree of inefficiency is thus inevitable when physicians vary in the value they place on patient health. By extension, the cost of resulting inefficiencies will be particularly large when the variation in practice styles is great. It is of interest, then, to consider this variation’s origins.

Researchers typically take the distribution of physician types as given. Physician concern for patient health distinguishes them from suppliers in most markets, however, making it worth inquiring as to how these preferences are formed. While this sense of duty may in part be intrinsic to the individual, it also reflects a core aspect of physician training. Ellis and McGuire’s model shows that there is value in developing this sense of duty as consistently as possible.\(^4\) Patients and insurance beneficiaries have reason to value both the strength

\(^3\)Recent work by Abaluck and Agha (2013), for example, investigates variation in physicians’ use of advanced diagnostic imaging in the context chest CT scans for pulmonary embolism.

\(^4\)One could demonstrate a proposition along the following lines: Let \(F(\cdot)\) be the cumulative distribution function describing the distribution of \(\theta\) across a population of physicians. Let \(G(\cdot)\) be a mean preserving spread of \(F(\cdot)\). Then for a given population of patients, \(G(\cdot)\) will be associated with lower patient welfare than \(F(\cdot)\) when insurers are constrained to select a single per-unit reimbursement rate.
of physician agency and the uniformity of its application. Evidence on the malleability of physicians’ practice styles (Molitor 2011) suggests that such efforts may be fruitful.

3 Implications of Physician Billing Practices

The model from Section 1 makes an important assumption about the quantity of care. In effect, it assumes that fee-for-service payments are based on uniformly verifiable quantities of care received by patients. Put another way, it assumes uniformity in physicians’ billing practices; the billed quantity of care and the care delivered to the patient are the same. In practice, it would be remarkable if this were actually the case.

Variation in the bills submitted for a given level of service provision need not reflect fraud. Medical billing is non-trivially complicated, and the resulting administrative costs are well documented (Cutler and Ly 2011). Variation in billing may reflect both differences in the resources devoted to the process and genuine uncertainty over appropriate billing procedures. Regarding the latter, Medicare periodically releases extensive guidance characterizing the appropriate billing of various scenarios (Center for Medicare and Medicaid Services 2013). The need for such guidance suggests uncertainty, and with it variation in existing billing practices (Kikano, Goodwin and Stange 2000). Even after guidance is issued, appropriate interpretations may remain open for debate. Additionally, documenting the justification for all billable procedures can be costly. Physicians may vary in their billing savvy, as well as in the administrative and managerial resources they choose to devote to the process.

The aggressiveness of physician billing can be modeled as a ratio describing the difference between the quantity of care received and the quantity billed. As before, let $q$ describe the quantity of care received. If the parameter $a$ describes the aggressiveness of billing, then

---

5Kikano et al. (2000) document extensive variation in the billing of Evaluation and Management services. They find that under-billing is as common as over-billing, suggesting that variation in coding practices should not be viewed predominantly as a reflection of bad faith behavior.
the quantity billed can be described as $q_b = a \times q$. The utility of a physician with billing parameter $a$ is thus

$$U = \theta[a + r \times a \times q - c(q)] + (1 - \theta)b(q).$$

(5)

Provision of an incremental service gives the physician additional revenue of $a \times r$, has a marginal cost of $MC(q)$, and provides an agency benefit of $(1 - \theta)MB(q)$. The physicians’ marginal utility from providing an additional unit of care is thus:

$$MU(q) = \frac{dU}{dq} = \theta[a \times r - MC(q)] + (1 - \theta)MB(q).$$

(6)

The physician maximizes his or her utility by choosing the quantity such that

$$\theta \times a \times r + (1 - \theta)MB(q) = \theta MC(q).$$

(7)

The optimal reimbursement rate can be found using the same sequence of steps applied in Section 1. The optimal reimbursement rate is

$$r^* = \frac{MB(q^*)}{a}[2 - \frac{1}{\theta}].$$

(8)

The solution has changed quite little from the solution found in Section 1. The optimal per-unit payment is simply scaled to account for the rate at which provided services are converted into billed services.

For any fixed level of the billing parameter, there is a per-unit payment that induces the efficient quantity of care provision. As in the previous section, problems arise when billing
savvy varies across physicians. Practices that bill aggressively will, in effect, receive higher per-unit payments for each service they provide. Consequently, a fixed payment regime will lead proficient billers to provide excessive quantities of care. Timid billers, who are effectively underpaid, will provide too little.\footnote{As in footnote 4, one could prove a proposition along the following lines: Let $F(\cdot)$ be the cumulative distribution function describing the distribution of $a$ across a population of physicians. Let $G(\cdot)$ be a mean preserving spread of $F(\cdot)$. Then for a given population of patients, $G(\cdot)$ will be associated with lower patient welfare than $F(\cdot)$ when insurers are constrained to select a single per-unit reimbursement rate.} Future research on variation in physician billing practices would enable a more thorough assessment of the cost of resulting inefficiencies.

4 Implications of the Cost Structures of Capital- and Labor-Intensive Services

This section discusses difficulties for Medicare’s payment model that arise due to differences in services’ cost structures. Medicare reimburses physicians on what is intended to be an average-cost basis. This has significant implications for the profit margins of capital-intensive services relative to labor-intensive services.

4.1 Low Fixed Costs: The Case of Standard Office Visits

First consider the case of reimbursement for office visits. Basic office visits earn physicians roughly $50 to $100, depending on the length of the visit and the complexity of the patient’s condition. In the language of Medicare’s fee schedule, physicians’ “own work,” accounts for roughly 50 percent of this payment. The remaining 50 percent involves “practice expenses.” The practice expenses associated with office visits primarily involve staff time and other variable costs. To a first approximation, one can think of office visits as services for which essentially all costs are variable costs. This contrasts sharply with the case of MRIs, which are analyzed below. In the latter case, roughly 80 percent of Medicare’s payment is associated
with the MRI machine itself, which involves a fixed-cost investment.

What are the implications of low fixed costs in systems of average-cost reimbursement? If marginal costs are constant, then the absence of fixed costs implies constant average cost. The reimbursement will thus equal the constant marginal and average cost. Costs associated with a physicians’ own time and with management of the practice are not literally constant. It is standard to assume that they are increasing, in particular as the practice pushes the time boundaries of the standard work week. Marginal costs will exceed the average-cost reimbursement once these sources of rising marginal costs kick in. As implied by equation (3), this will lead physicians to supply office visits to a point where the marginal benefit of care (net of patient cost-sharing and inconvenience) remains moderately positive.

4.2 High Fixed Costs: The Case of MRIs

Now consider the case of reimbursement for MRIs. Lowes (2013) explains:

For physicians performing diagnostic scans, one big expense is the cost of the machine, such as a $2 million MRI scanner. In 2009, the Medicare Payment Advisory Commission (MEDPAC) reported that the equipment needed for a 70553 MRI brain scan represented almost 90% of the TC fee.\(^7\)

CMS calculates the per-service cost of imaging equipment by multiplying how many minutes the machine is typically used for a particular scan by its cost per minute. The cost per minute equals the purchase price of the machine spread out over the number of minutes of likely use over its probable lifespan.

---

\(^7\)To unpack the jargon, “70553” refers to the billing code associated with an MRI of the brain that does not make use of contrast dye. “TC” refers to the technical component of the MRI, which is one of two distinct components of the coding associated with bills for MRI services. The technical component involves the actual administration of the MRI. The “professional component,” which is not mentioned above, refers to the interpretation of the resulting imagine. The professional component is the labor-intensive component of MRI provision. It accounts for roughly 10 percent of the total payment for a typical MRI service.
In recent years, the payment for a typical MRI service has been on the order of $1,000. The bulk of this payment is associated with “practice expenses,” which, as noted above, are tracked separately from the payment associated with physicians’ “own work.” As noted by Lowes (2013), the practice expense is associated primarily with the cost of the MRI machine itself.

Medicare’s methodology can be described roughly as follows. The payment for an MRI service is meant to equal average variable cost, $AVC$, plus average fixed cost, $AFC$. Average variable costs are estimated, as in the case of office visits, on the basis of required quantities of physician time, staff time, etc. How does Medicare estimate average fixed cost? Average fixed cost is the total cost of the machine, $F$, divided by Medicare’s estimate of utilization over the MRI machine’s lifetime, $q_{est}$. Medicare’s reimbursement for an MRI service can thus be written as

$$r = AVC + \frac{F}{q_{est}}.$$  

With a reimbursement of $1,000 and average variable costs of roughly $200, Medicare’s payment allows for an average fixed cost of $800. Given the $2 million cost of the MRI machine, the Center for Medicare and Medicaid Services (CMS) has assumed that the MRI machine would be operational for around 2,500 services.

How did CMS arrive at this estimate and how accurate was it in practice? Lowes (2013) writes that CMS based its initial calculation on the assumption that MRI machines would be in use for approximately 25 hours per week. This turned out to be a significant understatement. More recent evidence, which the Medicare Payment Advisory Commission (MEDPAC) has recommended adopting, suggests that actual usage is equivalent to 45 hours per week. By underestimateing utilization, CMS overstated the average fixed cost by 40 percent.
Recall equation (3), which characterized physicians’ choices regarding the quantity of care to provide. If variable costs were literally fixed at $200, then the net profit associated with MRI provision would be a constant $800. Equation (3) implies that physicians would provide MRIs until patients’ net benefits from the service are sufficiently negative to counteract this financial incentive. This need not be taken literally to mean that the physician will provide the service until it does harm. Negative net benefits may reflect low diagnostic benefits coupled with the patient’s cost-sharing and inconvenience. As above, the cost of the physicians’ own time, as well as the cost of managing the practice, would typically be assumed to rise as the scale of the practice increases. As supply increases, these rising marginal costs will moderately reduce the $800 in profit per service. It is nonetheless quite clear that average cost reimbursement will tend to result in an excess of MRI provision relative to office visits.

4.3 Payment Systems’ Influence on Entry and Investment

In the language of Ellis and McGuire’s payment model, Medicare Part B relies entirely on \( r \), the fee-for-service payment, and not at all on \( \alpha \), the salary-like payment. Indeed, the salary-like payment disappeared relatively early in the exposition of the model itself; optimal reimbursements have been described exclusively in terms of the per-unit reimbursement. What is the role of the salary-like payment and how might it be used to improve the reimbursement of capital-intensive services?

To this point, optimal payments have been described as payments that lead existing physicians to provide care until its marginal cost equals its marginal benefit. Importantly, health systems must also account for the development and entry of new physicians. By itself, the per-unit payment provides an incomplete characterization of the total return to practicing medicine. The total return, which determines the desirability of entering medical practice in the first place, depends on both components of the payment. Adjustments to
the salary-like payment can worsen or improve the attractiveness of entering into medical practice or making a variety of relevant investments. Margins of interest include the decision to become a physician, the decision of an existing practice to treat Medicare beneficiaries, and the decision to invest in imaging equipment like MRI machines.

Investment in MRI machines presents a classic case where mixing a salary-like payment with a fee-for-service, or per-use, payment makes sense. When payments only involve the per-unit reimbursement, Medicare faces a bind. Paying on the basis of average costs leaves substantial incentives for over-utilization. Marginal cost reimbursement, on the other hand, makes investment in imaging equipment unattractive. By financing a portion of the investment up front, Medicare can increase the attractiveness of investment without creating incentives for excess utilization.

4.4 The Evolution of Physicians’ Practices and the Development of New Technology

As discussed above, the decision to invest in advanced technology depends on the total return to doing so. Under average cost reimbursement, this hinges in large part on the accuracy of Medicare’s assessment of the machine’s lifetime usage. If Medicare significantly understates lifetime usage, as Lowes (2013) reports it did, the incentive to invest will be strong.

The history of physician investments suggests that the incentives created by Medicare had substantial effects. Total utilization of advanced imaging equipment rose substantially over recent decades. This was particularly true of utilization by physicians with self-referral opportunities (Mitchell 2007, Mitchell 2008, Baker 2010). Hospitals were similarly observed to respond to incentives favoring investments in capital equipment following Medicare Part A’s adoption of the Prospective Payment System (Acemoglu and Finkelstein 2008).
The incentives Medicare created, and the investments that ensued, have two additional long-run implications. First, investment in advanced imaging technology alters the cost-structure of physicians’ practices (Clemens and Gottlieb 2013). Such investments make physicians’ practices more capital intensive, which widens the spread between marginal and average costs. Over time, these investments have thus shifted physicians away from the types of services for which Medicare’s fee schedule functions best. Adapting the fee schedule to changes in the environment has thus become an increasingly urgent task.

Finally, physicians’ investments shape the markets for innovation in advanced imaging equipment. When fee schedules encourage physicians to expand their provision of diagnostic services, they create markets for advances in these technologies. A growing body of research shows that the incentives created by factors including demographics, health policy, and insurance arrangements have substantial effects on innovative activity. Such evidence has been found in the context of vaccine development (Finkelstein 2004), in the broader pharmaceutical context (Acemoglu and Linn 2004), in the development of cancer treatments (Budish et al. 2013), and in the development of medical equipment and devices (Clemens 2012). In addition to shaping current care access and health expenditures, Medicare’s fee schedule shapes the treatments and tools that will be at physicians’ disposal in future years.
References


Godager, Geir, “Profit or Patients Health Benefit? Exploring the Heterogeneity in Physician Altruism,” *Journal of Health Economics*, 2013, (0), –.


