

AN ECONOMIC EXPLANATION FOR FRAUD AND ABUSE IN PUBLIC MEDICAL CARE PROGRAMS

ROGER FELDMAN*

ABSTRACT

This paper comments on David Hyman's theory of fraud and abuse in medical care. It agrees with Hyman that preventing fraud is difficult because providers, patients, and program administrators usually have weak incentives to do so. It extends Hyman's work by arguing that the root cause of fraud in public-sector medical programs is distorted prices (usually too high), coupled with limitations on efficiency-seeking activities that normally would occur when prices are distorted. The theory is illustrated with examples from Medicare, kickbacks and fee splitting, and a model of the behavior of fraud-control officers.

EVERYONE is against fraud and abuse—until they actually encounter it. As David Hyman's informative paper illustrates, addressing the problem of fraud and abuse is more difficult than condemning it.

The problem arises in part because an act of medical fraud usually involves four parties, either directly or indirectly. The first party is the provider, who regards his or her primary interest to be fidelity to patients. This is especially true of physicians, who are trained not only to represent patients but also to resist the "interference" of nonmedical personnel in regulating their professional conduct. Hyman reports that a particular type of fraud—manipulation of reimbursement rules—is seen by a sizeable percentage of physicians as necessary to maintain high-quality care for patients.

Patients are the second party involved in medical fraud. Hyman points out that patients are of two minds about fraud. Although they disapprove strongly of raw fraud, their feelings change as soon as the issue involves their own medical care. I am sure that many readers of this paper know of a doctor who has upcoded a visit to secure third-party reimbursement to avoid the prying eyes of a managed care reviewer. I would guess that most patients do not want to know what goes on inside the "black box" of medical reimbursement policy, but if they did, they would support this type of fraud on the part of physicians if it affected their own medical care.

* Blue Cross Professor of Health Insurance, University of Minnesota. This paper was prepared as a comment on David A. Hyman, *Health Care Fraud and Abuse: Market Change, Social Norms, and the Trust "Reposed in the Workmen,"* in this issue, at 531.

[*Journal of Legal Studies*, vol. XXX (June 2001)]

© 2001 by The University of Chicago. All rights reserved. 0047-2530/2001/3002-0012\$01.50

Third, program administrators may have an interest in looking the other way when fraud occurs. For example, they may be willing to overlook questionable behavior by a physician who serves a poor neighborhood or a hospital with a disproportionate share of Medicaid patients. Aggressive prosecution of these providers might reduce the supply of services that are judged to have positive externalities. Program administrators also tend to adopt “pro-provider” attitudes, especially in public programs where the emphasis historically has been on encouraging provider participation.

It seems that the only party with an interest in controlling fraud is the fraud-control officer. No wonder he is perceived as the grinch in this story.

The power of social norms can be illustrated by an experience of a colleague of mine who, for his doctoral dissertation, studied a program that brought nurse practitioners (NPs) into remote rural areas of Appalachia. While documenting the services these NPs performed, my colleague kept coming across medical encounter forms for “Dog Jones” and “Cat Smith.” Why did the locals prefer these odd names? As he pursued the answer, he found that Dog Jones was Mr. Jones’s dog and Cat Smith was Ms. Smith’s cat. The NPs were treating sick pets along with their owners and were dutifully filling out the paperwork. Was this a fraud? The private foundations and government agencies that funded the NP program may have thought so (although it is not clear that they would have jeopardized the program by attacking its by-product pet care), but the providers and patients clearly did not see anything wrong with treating pets.

Nevertheless, while social norms may provide the grease that makes fraud so slippery to catch, I do not think they explain why it occurs so often in public medical care programs. To really understand the origins of medical fraud, we need to look at the economic incentives that lie behind provider and consumer behavior. More specifically, I will argue that the root cause of public sector medical fraud is distorted prices, coupled with limitations on efficiency-seeking economic activities that normally would occur when prices are distorted.

I will illustrate this theory with three examples. The first example is taken from the Medicare program. Since the early 1980s, Medicare beneficiaries have had a choice of two medical care systems. First, there is traditional “fee for service” (FFS), which was the basis for the original Medicare entitlement. Under the FFS system, the beneficiary may obtain covered services from any qualified and willing provider, who receives a fee for each service. The other system is organized around risk-bearing “Medicare+Choice” (M+C) health plans that receive a prospective payment from the government on a monthly basis for each person who enrolls in the plan. Until 1997, the monthly payment was tied directly to the FFS cost of the entitlement benefit package in the beneficiary’s county. The Balanced Budget Act of 1997¹ altered the

¹ Pub. L. No. 105-33 (1997).

direct link between FFS costs and payments to M+C plans. Very high and low payment areas were compressed toward the national average. But the primary determinant of payment remains the historical cost of FFS Medicare in the county.

The current payment system has resulted in large disparities in payments to M+C plans because FFS Medicare costs vary widely among counties. In 2000, for example, M+C plans in Miami received \$794.02 per month for each “standard” beneficiary who enrolled, while plans in Minneapolis received only \$457.66 per month. There are similar wide differences in payments among other counties around the 2000 national average rate of \$504.96 per month. These variations are much larger than differences in the M+C plans’ costs, which are more uniform nationally.²

In high-payment areas, M+C plans may offer enhanced benefits—such as outpatient prescription drugs—that exceed the Medicare entitlement. However, they are not allowed to convert any of the excess government payment into premium rebates (that is, they may not write checks to enrollees). I have suggested elsewhere³ that it is inefficient to prevent M+C plans from giving premium rebates because some of the enhanced benefits currently offered by plans in high-payment areas are not worth as much to beneficiaries as they cost to produce. Given the choice, enrollees in these plans would rather have a cash rebate than extra benefits.

Since 1995, I have been part of a team of advisers that has developed a demonstration design for replacing the current method of paying M+C plans with one more closely based on the plans’ costs.⁴ The design would use competitive bids submitted by the plans to set the government payment rate in each demonstration market. As part of this effort, we recommended that low-bidding plans be allowed to give premium rebates. If a low-bidding M+C plan offered a premium rebate, it would reduce the price of that plan below the Part B premium that FFS beneficiaries would still have to pay for their coverage, thereby creating direct price competition for the first time between M+C plans and FFS Medicare.

Although premium rebates represented an opportunity to inject competitive pricing into Medicare, the Centers for Medicare and Medicaid Services (CMS) (formerly known as the Health Care Financing Administration) demurred, on the grounds that rebates would create substantial administrative difficulties. Furthermore, the Department of Health and Human Services Office of the General Counsel (OGC) warned that rebates would present legal

² Roger Feldman *et al.*, *An Empirical Test of Competition in the Medicare HMO Market, in Competitive Approaches to Health Policy Reform* (R. J. Arnould, R. F. Rich, & W. D. White eds. 1993).

³ Roger Feldman *et al.*, *Premium Rebates and the Quiet Consensus of Market Reform for Medicare*, 23 *Health Care Financing Rev.* 19 (winter 2001).

⁴ Bryan Dowd, Robert Coulam, & Roger Feldman, *A Tale of Four Cities: Medicare Reform and Competitive Pricing*, 19 *Health Aff.* 9 (2000).

difficulties because they would conflict with the antikickback statutes. It was not until Clinton administration offered its own Medicare reform proposal that included rebates that the OGC relented and decided that providing rebates up to the Part B premium would be legal under the demonstration.

This example illustrates the complexity of medical kickbacks and how the government can overlook a proposed arrangement that may violate the law but appears consistent with other goals (in this case, the administration's own reform proposal). But the fundamental problem here is not a legal one. The problem is that the FFS-based payment rate for M+C plans is too high in many market areas. Marketplace competition would tend to convert this overpayment into cash—a move that would promote economic efficiency. But the Medicare program does not allow this efficiency-improving measure, and therefore competition is forced to take place over inefficient benefit enhancements such as free health club memberships and free transportation to the doctor's office.

In addition to showing how a ban on premium rebates reduces economic efficiency, this example illustrates another point. Suppose we ask what type of fraud is likely to occur under a prospective payment system for Medicare health plans. A naive prediction would be that plans will underserve their members. That is, their incentive is to "take the money and run." This prediction is not necessarily correct. If the government payment is set above the level that would cover all benefits that enrollees are willing to pay for and competition is present, the result is likely to be overprovision of services to M+C enrollees.

My second example also relates to the economics of kickbacks and fee splitting. In 1979, Mark Pauly proposed an economic model of kickbacks that still seems accurate, both in its overall approach and in the particular conclusions it reached.⁵ Pauly imagined that a patient initially consults a "generalist" physician who must decide what services to recommend for the patient, including services that may be performed by the generalist or by "specialists" to whom the patient may be referred.

If every activity performed by both types of physicians were priced at its minimum marginal cost, including the cost of physician time, kickbacks would be irrelevant. No provider would be willing to offer a kickback, and none would be willing to pay. In order to have a kickback, we need some excess of price over marginal cost. Continuing with this line of reasoning, Pauly concluded that either or both of two conditions must be present: there may be monopoly in the market for potentially referred services, whether performed by the generalist or the specialist, or "there may be public or private insurers who mistakenly set the fee for the activities too high."⁶

⁵ Mark V. Pauly, *The Ethics and Economics of Kickbacks and Fee Splitting*, 10 *Bell J. Econ.* 344 (1979).

⁶ *Id.* at 347.

The equilibrium in this market will be determined as follows. As long as price less the prevailing level of kickbacks is greater than the specialist's marginal cost, the specialist can increase his profit by offering a larger kickback. Equilibrium therefore requires that the specialist's fee less the kickback equals his marginal cost. Whether the generalist decides to refer or not depends on the kickback offered by the specialist relative to the net income the generalist could earn by performing the service himself. There will be some set of recommendations that are acceptable to patients and for which referral with kickbacks is compatible with the physicians' incentives. Furthermore, banning kickbacks for those activities would not prevent the services from being delivered but would simply lead generalists to perform them "in-house," at a higher marginal cost to society. As in the case of a ban on premium rebates, banning kickbacks in the presence of distorted fees prevents physicians from improving the efficiency of their care for patients.

Pauly's theory can be illustrated by the abuse of the durable medical equipment, prosthetics, orthotics, and supplies (DMEPOS) benefit in Medicare. Prices for DMEPOS are widely thought to be excessively high. High prices result in large profits that make it possible to share returns among multiple users in the form of kickbacks, sham intermediaries, and other profit-splitting mechanisms. According to a report by Abt Associates,⁷ the fundamental problem is "that Medicare prices bear no relationship to the underlying economic cost of the DMEPOS." The report concluded that "Medicare DMEPOS payment levels appear sufficiently high to constitute the largest unnecessary cost of the DMEPOS benefit, possibly exceeding all other forms of abuse."

My last example concerns the economics of fraud control itself. I will analyze the incentives of fraud-control officers (FCOs) who are paid a fee for each fraudulent service they detect. If, as Hyman suggests, the FCOs currently supply too much control, this implies that the fee is too high and should be reduced.

In this model, the FCO pursues "utility" or satisfaction that is directly related to the dollar volume of fraud detected and inversely related to the level of "effort" it takes to catch the fraud. This is admittedly a simple view of the FCO's objectives that makes the FCO seem like a bounty hunter, but I think it is not far from the mark. For example, the CMS uses the dollar volume of denied claims as one measure for evaluating the performance of the private carriers it hires to pay Medicare physician claims. Since February 1998, Medicare has encouraged beneficiaries to get into the act by offering them a reward of up to 10 percent of the Medicare funds recovered when

⁷ Robert Coulam & Leo Reardon, *Models of Abuse of the Medicare DMEPOS Benefit* 23 (Abt Associates Report, September 1993).

they turn in providers who engage in fraud and abuse against the Medicare program.⁸

Furthermore, a variety of external forces have increased the FCO's incentive to measure success by the volume of fraud it detects. One of these is the presence of *qui tam* (whistleblower) actions under the False Claims Act. Congress amended the act in 1986 to increase the power of relators to bring lawsuits and increased the whistleblower's share of the recovery. Since 1988, nearly \$2 billion have been recovered from health care providers and others who have cheated government health programs.⁹ Prior to 1996, the whistleblower got 15–25 percent of the total recovery with the balance going to the Treasury (all the money went to the Treasury in cases where the government initiated the lawsuit). The passage of the Health Insurance Portability and Accountability Act in 1996¹⁰ changed these structural arrangements, so that the government's share now goes straight to the Medicare trust fund.¹¹ Although this is not a pure bounty system, it is much closer than had previously been the case. In fiscal year 1999 alone, \$114.4 million was deposited in the Medicare trust fund, largely from penalties, damages, and criminal fines.

Another factor motivating the FCO to view its mission as recovering fraudulent payments is harsh reviews from Congress and government oversight agencies.¹² Since 1997, Congress has required the Departments of Health and Human Services and Justice to issue joint annual reports for the preceding year on the amounts of fraud and abuse recovered. With these structural incentives all motivating the FCO to recover fraudulent claims, the following model takes on a reasonable degree of plausibility.

Let the amount of resources expended by the FCO to detect fraud be denoted by E , which I will refer to as “effort.” I assume that the amount of fraud detected is equal to the product of effort times the number of “fraudulent acts” committed (F). A fraudulent act might be the submission of a claim for services that were not performed. Because providers are aware of the penalties for getting caught, they will submit fewer fraudulent claims as the

⁸ This bounty—dubbed the Incentive Reward program—is described on Medicare's Web site (<http://www.medicare.gov/fraudabuse/overview.asp>) and in various CMS publications.

⁹ Shelly R. Slade, *Health Care Fraud: How Far Does the False Claims Act Reach?* (August 23, 2000) (<http://www.quackwatch.com/02ConsumerProtection/fca.html>, visited August 6, 2001).

¹⁰ Pub. L. No. 104-191 (1996).

¹¹ David A. Hyman, *HIPAA and Health Care Fraud: Where's the Beef?* 22 *Cato J.* (forthcoming 2002).

¹² U.S. General Accounting Office, *Health Insurance: More Resources Needed to Combat Fraud and Abuse* (GAO Testimony T-HRD-92-49, July 28, 1992); *Medicare: Concerns about HCFA's Efforts to Prevent Fraud by Third-Party Billers* (GAO Testimony T-HEHS-00-93, April 6, 2000); *Medicare: Health Care Fraud and Abuse Control Program Financial Reports for Fiscal Years 1998 and 1999* (Report No. GAO/AIMD-00-275R HCFAC 1998 and 1999, July 31, 2000).

FCO's detection effort increases. The number of fraudulent acts committed can be expressed as $F = F(E)$, where $F' < 0$ (the prime symbol denotes a derivative, or the change in fraud with respect to a small change in FCO effort). Finally, I assume that effort is a scarce commodity, and so, other things being equal, the FCO would rather spend less of it. The disutility of effort is much like the disutility of work that is assumed by neoclassical labor economists.¹³ This factor can be summarized by the function Ψ , which stands for the disutility of effort, with $\Psi' > 0$ and $\Psi'' > 0$. With these assumptions, the FCO's utility function can be written as

$$U = PEF - \Psi(E), \quad (1)$$

where P is the reward or "price" per unit of fraud detected. We do not need to assume that the FCO literally receives this fee; it is sufficient to argue (as I did above) that FCO utility is proportional to the amount of fraud detected, with P being the factor of proportionality.

The first-order condition for maximizing equation (1) is

$$U_E = P(EF' + F) - \Psi' = 0, \quad (2)$$

where U_E is the first derivative of utility with respect to effort. In equilibrium, the marginal return from pursuing fraud will equal the marginal disutility of effort.

Next, I address the question of what happens when P changes. I assume that P is controlled "exogenously" (that is, outside the influence of the FCO) by the program administrator, which is the CMS in the case of Medicare. The answer to this question involves totally differentiating equation (2) and solving for dE/dP :

$$dE/dP = -(EF' + F)/U_{EE}, \quad (3)$$

which is positive if the second-order condition for utility maximization is satisfied. In other words, the FCO's effort is an increasing function of the fee it receives. By controlling the fee, the CMS can determine the amount of effort expended and thus both the amount of fraud committed and the amount detected.

The CMS could pursue any one of numerous objectives in setting the fee. One of these is to minimize the sum of the cost of fraud committed and FCO effort, or $S = F + \Psi(E)$. The first-order condition for minimizing this objective function is

$$S_E = F' - \Psi' = 0, \quad (4)$$

which says that the marginal product of effort in deterring fraud should equal the marginal disutility of effort for the fee to be socially optimal. Looking

¹³ Albert Rees, *The Economics of Work and Pay* 22 (1973).

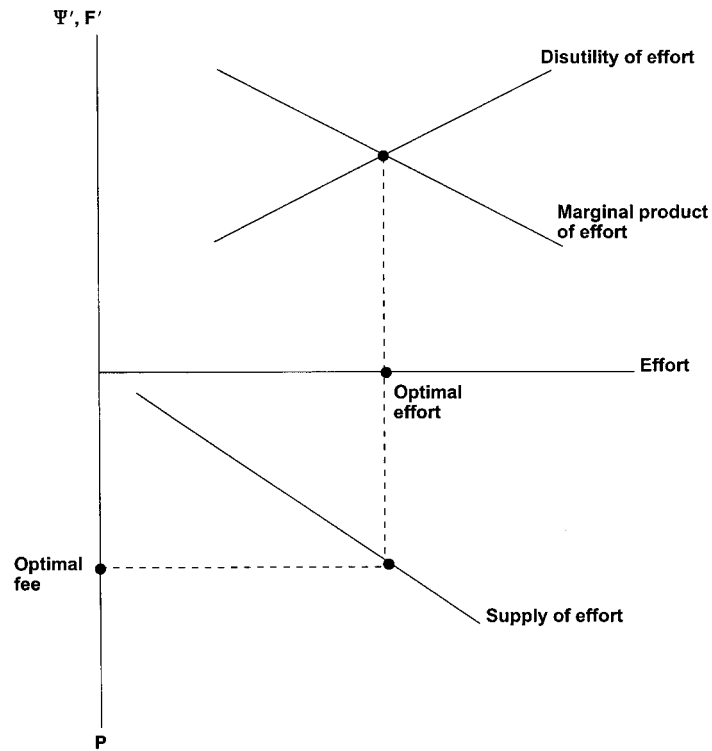


FIGURE 1.—Optimal fraud control effort and pricing

back at the FCO's maximization problem and substituting $F' = \Psi'$ into (2), the CMS should set $P = F'/(EF' + F)$ to minimize the social cost of fraud.

The concepts of this model are illustrated in Figure 1, where the upper panel shows the marginal product of effort and the marginal disutility of effort. The optimal amount of effort is determined where these functions intersect. The supply of effort is shown in the lower panel of Figure 1. Here the CMS calculates the fee that is required for the FCO to supply the optimal amount of effort. If it is generally felt that the current supply of effort (and thus the supply of fraud control) is too great, the inescapable conclusion is that the fee is too high.

In summary, Hyman draws attention to the social norms that make the problem of correcting fraud and abuse extremely difficult. I have attempted to supplement his argument with a discussion that highlights the problem of inappropriate incentives for medical care providers and consumers. Usually,

these incentives point toward prices that are maintained above the competitive level, coupled with restrictions on economic activities (such as fee splitting and premium rebates) that would allow some of the providers' bottled-up profit-seeking tendencies to be realized.